VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Vážené varianty algoritmů pro určení shlukovacího koeficientu v komplexních sítích

Algorithms for Clustering Coefficient in Weighted Complex Networks

2019 Petr Žoček

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Zadání bakalářské práce

Student: **Petr Žoček**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 2612R025 Informatika a výpočetní technika

Téma: Vážené varianty algoritmů pro určení shlukovacího koeficientu v komplexních sítích
Algorithms for Clustering Coefficient in Weighted Complex Networks

Jazyk vypracování: čeština

Zásady pro vypracování:

Zkoumání struktury a procesů v komplexních sítích reprezentujících reálné sítě různých typů (technické, informační, sociální apod.) je v současnosti dynamicky se vyvíjející oblastí. Určování shlukovacího koeficientu je jedním z častých testů pro analýzu komplexních sítí. Cílem práce je implementace algoritmů pro výpočet shlukovacího koeficientu pro ohodnocené (vážené) komplexní sítě.

1. Seznamte se s komplexními sítěmi a s vlastnostmi, které se nejčastěji zkoumají.
2. Seznamte se se shlukovacím koeficientem a se způsobem jeho určení pro ohodnocené i neohodnocené (vážené i neválené) grafy.
3. Vyberte a naimplementujte algoritmy pro určení shlukovacího koeficientu.
4. Nad vybranými kolekcemi dat proveďte experimenty a jejich výsledky vhodně reprezentujte.

Seznam doporučené odborné literatury:

[1] M. E. J. Newman, Networks: An Introduction, Oxford University Press (2010), ISBN-10: 0199206651.
[2] http://toreopsahl.com/
[3] Dle pokynů vedoucího bakalářské práce.

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **RNDr. Eliška Ochodková, Ph.D.**

Datum zadání: 01.09.2017
Datum odevzdání: 30.04.2019

doc. Ing. Jan Platoš, Ph.D.
vedoucí katedry

prof. Ing. Pavel Brandštetter, CSc.
děkan fakulty

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě 30. dubna 2019 ..... Švec

Rád bych na tomto místě poděkoval vedoucí práce RNDr. Elišce Ochodkové, Ph.D. za odborné vedení, věcné konzultace a rady pro vypracování této práce.

Abstrakt

Tato práce se zabývá výpočty shlukovacího koeficientu pro ohodnocené komplexní sítě. Cílem práce je implementace algoritmů těchto výpočtů. Konkrétně čtyř variant, které ve svých pracích navrhli Barrat, Onnela, Zhang a Holme. Součástí práce je výběr vhodné datové struktury pro reprezentaci sítě a vizualizace výsledků výpočtů. Výsledkem práce jsou programové komponenty obsahující vytvořené implementace, rozšířené o jednoduché grafické rozhraní pro práci daty. Součástí výsledků je sada experimentů provedených a zpracovaných pomocí vytvořeného rozhraní.

Klíčová slova: komplexní sítě, shlukovací koeficient

Abstract

This work deals with the calculation of the clustering coefficient in weighted complex networks. The goal of this work is to implement algorithms of these calculations. Specifically, the four variations proposed by Barrat, Onnela, Zhang and Holme in their works. Part of the work is selection of suitable data structure for network representation and visualization of calculated results. The result of this work are program components containing created implementations, extended by simple graphical interface for data processing. The results include a set of experiments performed and processed by the created interface.

Key Words: complex networks, clustering coefficient

Obsah

Seznam použitých zkratek a symbolů

CSV	–	Comma Separated Values
GUI	–	Graphical User Interface
ISP	–	Internet Service Provider
PC	–	Personal computer
PNG	–	Portable Network Graphics
USA	–	United States of America
WWW	–	World Wide Web

Seznam obrázků

Seznam tabulek

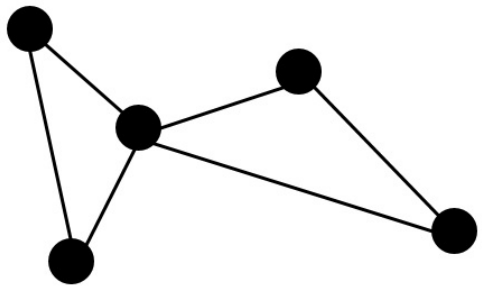
1 Úvod

Věda o sítích je nová disciplína, která vznikla v první dekádě 21. století. Zabývá se zkoumáním struktury a procesů v komplexních sítích reprezentujících reálné sítě různých typů (technologické, informační, sociální a biologické). V současnosti jde o dynamicky se vyvíjející oblast. Určování shlukovacího koeficientu je jedním z častých testů pro zkoumání vlastností komplexních sítí. Cílem této práce je vytvořit implementace výpočtů pro určení shlukovacího koeficientu ohodnocených komplexních sítí a následně je využít pro experimenty nad různými datovými sadami reprezentujícími jednotlivé typy sítí.

Ve druhé kapitole se seznámíme se základy teorie grafů, které jsou nutným předpokladem pro pochopení následujících kapitol. Ve třetí kapitole se seznámíme s komplexními sítěmi. Povíme si, proč komplexní sítě studujeme a jaké vlastnosti zkoumáme. Ukážeme si jednotlivé typy komplexních sití, rozdíly mezi nimi a jejich příklady v realném světě okolo nás. Ve čtvrté kapitole se seznámíme se shlukovacím koeficientem. Rozebereme vážené (ohodnocené) varianty výpočtu a rozdíly mezi nimi. V páté kapitole přistoupíme k implementaci těchto výpočtů. Vytvoříme knihovnu výpočtů a jako podklad provádění experimentů bude vytvořeno jednoduché grafické uživatelské rozhraní. V šesté kapitole budou provedeny experimenty s implementovanými výpočty nad vybranými datovými sadami. Výsledky budou zaznamenány a vhodně prezentovány. V sedmé kapitole budou zhodnoceny výsledky této práce.

2 Základy teorie grafů

Pro správné pochopení následujících kapitol, je pořeba seznámit se se základy teorie grafů. Ta je důležitou součástí diskrétní matematiky, která pomáhá při řešení praktických úloh. Využitím teorie grafů modelujeme reálnou úlohu jako množinu objektů (vrcholy grafu) a vztahů mezi nimi (hrany grafu). Nespornou výhodou je snadná implementace objektů a postupů teorie grafů v počítači. Hlavním zdrojem při psaní této kapitoly byla skripta [2][3].



Obrázek 1: Ukázka grafu

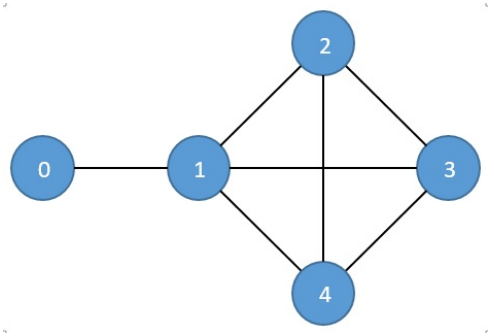
2.1 Graf

Graf je ústředním pojmem teorie grafů. Jedná se o model, který reprezentuje objekty a vztahy mezi nimi. Graf značíme symbolem G. Ve své nejjednodušší formě si jej můžeme představit jako puntíky a čáry mezi nimi, viz. obrázek 1. Puntíky reprezentují vrcholy grafu. Vrcholy grafu budeme značit v[1], dále v textu je budeme značit indexem. Množinu vrcholů grafu budeme značit $V(G)$[2], pokud nebude hrozit omyl použijeme zjednodušené označení V. Počet vrcholů v grafu budeme značít n, lze se však setkat s označením $|V(G)|$. Čáry reprezentují hrany grafu. Hrany budeme značit e[3]. Každá hrana má dva koncové vrcholy, se kterými je incidentní. Incidencí nazýváme vztah mezi hranou a jejími koncovými vrcholy. Je-li vrchol v koncovým vrcholem hrany e, můžeme psát $v \in e$ a říkáme, že vrchol v je incidentní s hranou e nebo také že hrana e je incidentní s vrcholem v. Množinu hran grafu budeme značit $E(G)$[4], stejně jako u množiny vrcholů i zde lze použít zjednodušené označení E. Počet hran v grafu značíme m, lze se však setkat s označením $|E(G)|$.

2.1.1 Jednoduchý graf

Základním typem grafu je jednoduchý graf. Příklad jednoduchého grafu můžeme vidět na obrázku 2.

[1] Z anglického vertex (vrchol)
[2] Z anglického vertices (vrcholy)
[3] Z anglického edge (hrana)
[4] Z anglického edges (hrany)



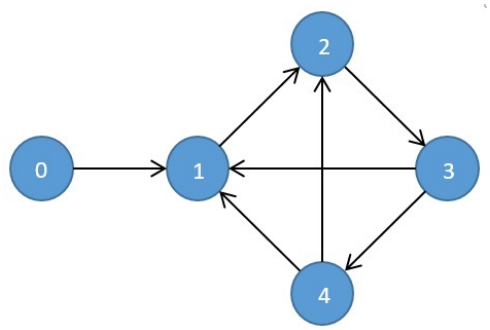
Obrázek 2: Jednoduchý graf

Definice 1 *Jednoduchý graf G je uspořádaná dvojice (V, E), kde V je neprázdná množina vrcholů a E je nějaká množina dvouprvkových podmnožin množiny V. Prvkům E říkáme hrany [3].*

Hrany jednoduchého grafu jsou tedy neuspořádané dvojice vrcholů $e_{ij} = \{v_i, v_j\}$. Takové hrany nazýváme neorientovanými hranami. Graf, který obsahuje pouze neorientované hrany, nazýváme neorientovaný graf. Jednoduchý graf je tedy neorientovaný graf. Dále z definice vyplývá, že jednoduchý graf nedovoluje existenci smyček (hran, které mají oba koncové vrcholy stejné), protože taková hrana by nebyla dvouprvkovou podmnožinou V. Stejně tak tato definice nedovoluje existenci multihran (více než jedné hrany mezi dvěma vrcholy), což vychází z předpokladu, že v množině se může vyskytovat každý prvek maximálně jednou.

2.1.2 Orientovaný graf

Často se můžeme setkat se situací, kdy mezi objekty skutečného systému není oboustranná vazba. Pro lepší zachycení takových situací byl zaveden koncept orientovaného grafu.



Obrázek 3: Orientovaný graf

Definice 2 *Orientovaným grafem rozumíme uspořádanou dvojici $G = (V, E)$, kde V je množina vrcholů a množina orientovaných hran je $E \subseteq V \times V$ [2].*

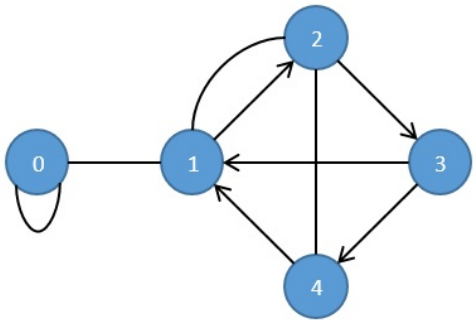
Orientovaná hrana pak již není dvouprvková podmnožina V, ale uspořádaná dvojice vrcholů $e_{ij} = (v_i, v_j)$. První vrchol z této uspořádané dvojice nazýváme počáteční, druhý pak koncový.

Na obrázku 3 si můžeme všimnout, že orientované hrany znázorníme šipkou u koncového vrcholu hrany. Všimněme si, že definice orientovaného grafu také nedovoluje existenci multihran. Povoluje však existenci smyček, což pro nás nemusí být vždy žádoucí. Proto si zde zavedeme ještě další pojem:

Definice 3 *Orientovaný graf, ve kterém navíc nejsou smyčky (ani násobné orientované hrany), bývá v literatuře označován jako jednoduchý orientovaný graf, případně jednoduchý digraf [3].*

2.1.3 Obecný graf

Poslední varianta grafu, se kterou se seznámíme, je obecný graf. To je graf, ve kterém mohou existovat orientované i neorientované hrany, smyčky a multihrany. Příklad obecného grafu můžeme vidět na obrázku 4. Tento typ grafu nám umožňuje co nejpodrobněji zachytit i velmi složité reálné situace.



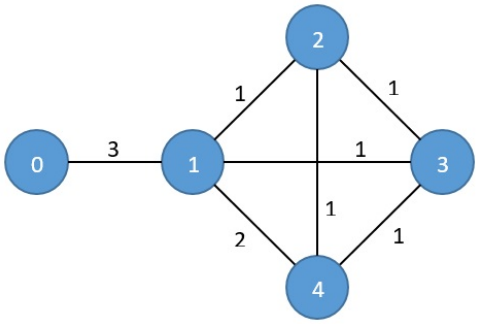
Obrázek 4: Obecný graf

Definice 4 *Obecný graf je trojice* (V, E, φ)*, kde* V *je neprázdná množina vrcholů,* E *je množina hran,* $E \cap V = \emptyset$*, a* φ *je incidenční zobrazení* $\varphi : E \to \binom{V}{2} \cup V^2 \cup V$ *[3].*

Všimněme si, že v této definici E představuje vlastně multimnožinu, která narozdíl od množiny umožňuje vícenásobný výskyt stejného prvku.

2.2 Ohodnocený graf

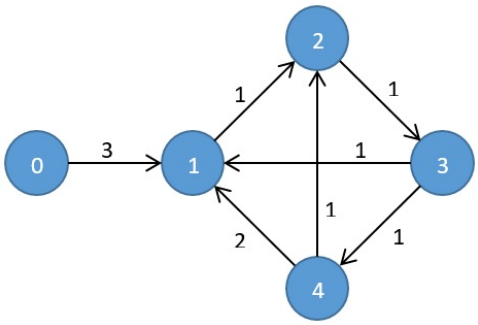
Při převádění reálných situací na grafy, můžeme narazit na situaci, kdy si nevystačíme s binárními hranami (hrana existuje - ”1”, nebo neexistuje - ”0”). Proto byl vytvořen koncept ohodnocených hran, kdy je každé hraně v grafu přiřazeno reálné číslo.



Obrázek 5: Ohodnocený jednoduchý graf

Definice 5 *Ohodnocení grafu G je funkce w: $E(G) \to \mathbb{R}$, která každé hraně $e \subseteq E(G)$ přiřadí reálné číslo $w(e)$, kterému říkáme váha hrany (značení w pochází z anglického „weight“). Ohodnocený graf je graf G spolu s ohodnocením hran reálnými čísly. Kladně ohodnocený (říkáme také vážený) graf G má takové ohodnocení w, že pro každou hranu $e \subseteq E(G)$ je její váha $w(e)$ kladná [2].*

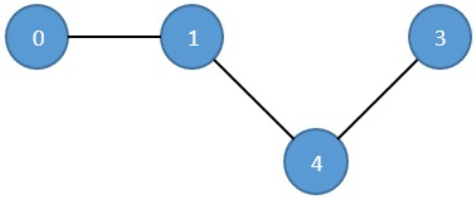
Ohodnocené mohou být všechny 3 typy grafů (jednoduchý, orientovaný a obecný), se kterými jsme se seznámili. Na obrázku 5 můžeme vidět příklad ohodnoceného jednoduchého grafu. Váha hrany je obvykle v grafu zapsána jako číslo vedle hrany, ale může být i vyjádřena jinými způsoby, například tloušťkou nebo barvou hrany. Na obrázku 6 můžeme vidět příklad ohodnoceného orientovaného grafu.



Obrázek 6: Ohodnocený orientovaný graf

2.3 Podgraf

Při řešení úloh se často setkáváme s případy, kdy z daného grafu vynecháme některé vrcholy (a hrany s nimi incidentní) nebo hrany, případně obojí. Takto vzniklý graf nazýváme podgrafem. Na obrázku 7 můžeme vidět příklad podgrafu vytvořeného z jednoduchého grafu na obrázku 2.



Obrázek 7: Podgraf

Definice 6 *Graf H nazveme podgrafem grafu G, jestliže $V(H) \subseteq V(G)$ a $E(H) \subseteq E(G)$, píšeme $H \subseteq G$ [2].*

2.3.1 Indukovaný podgraf

Speciálním typem podgrafu je indukovaný podraf.

Definice 7 *Podgraf I grafu G nazveme indukovaným podgrafem grafu G, jestliže $E(I)$ obsahuje všechny hrany grafu G, které jsou incidentní s vrcholy z $V(I)$. (Vynecháme pouze hrany, které byly incidentní s vynechanými vrcholy.) [2]*

2.3.2 Okolí vrcholu

Jsou-li dva vrcholy spojeny hranou, říkáme, že spolou sousedí. Množinu všech vrcholů, které sousedí s vrcholem i nazýváme okolí vrcholu i.

Definice 8 *Okolí vrcholu i je množina $N_i(G) = \{v_j \in V : \exists e_{ij} \in E(G)\}$ [3].*

Pokud nehrozí nedorozumění můžeme místo $N_i(G)$[5] použít zjednodušený zápis N_i.

2.4 Stupeň vrcholu

Stupeň vrcholu i budeme v této práci značit k_i. V neorientovaných grafech je definovan následovně:

Definice 9 *Stupeň vrcholu i v grafu G je definován jako počet hran, se kterými je vrchol i incidentní [2].*

V orientovaných grafech pak rozlišujeme vstupní stupěň vrcholu k_i^+, který vyjadřuje počet příchozích hran do vrcholu i, a výstupní stupěň vrcholu k_i^-, který vyjadřuje počet odchozích hran z vrcholu i.

[5] Z anglického neigbourhood (sousedství)

2.5 Reprezentace grafů

Jednou se zmíněných výhod použití grafu je jejich snadná implementace v počítačích. V následujících řádcích si představíme základní možnosti repretentace grafů, jejich výhody, nevýhody a použití.

2.5.1 Matice sousednosti

Prvním způsobem reprezentace grafů je matice sousednosti, často označována jako A. Pokud A reprezentuje neohodnocený jednoduchý graf:

$$A_{ij} = \begin{cases} 1 & \text{pokud } i \text{ a } j \text{ jsou spojeny hranou} \\ 0 & \text{jinak} \end{cases}$$

Matice sousednosti jednoduchého grafu na obrázku 2 vypadá následovně:

$$A_{i,j} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Můžeme si všimnout, že matice sousednosti jednoduchých grafů je symetrická. Matici sousednosti orientovaného grafu označujeme také A a je zpravidla nesymetrická. Při zápisu hrany rozlišujeme počáteční a koncový vrchol hrany. Řádky matice reprezentují index počátečního vrcholu hrany a sloupce index koncového vrcholu hrany. Takto zapsaná matice sousednosti orientovaného grafu na obrázku 3 pak vypadá následovně:

$$A_{i,j} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

Pokud A reprezentuje ohodnocený graf pak:

$$A_{ij} = \begin{cases} w_{ij} & \text{pokud } i \text{ a } j \text{ jsou spojeny hranou} \\ 0 & \text{jinak} \end{cases}$$

kde w_{ij} reprezentuje váhu hrany mezi vrcholy i a j. Matice sousednosti ohodnoceného jednodu-

chého grafu na obrázku 5 vypadá následovně:

$$A_{i,j} = \begin{pmatrix} 0 & 3 & 0 & 0 & 0 \\ 3 & 0 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 2 & 1 & 1 & 0 \end{pmatrix}$$

Matice sousednosti ohodnoceného orientovaného grafu na obrázku 6 vypadá následovně:

$$A_{i,j} = \begin{pmatrix} 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 2 & 1 & 0 & 0 \end{pmatrix}$$

Maticí sousednosti můžeme tedy reprezentovat jednoduché grafy, které mohou být také ohodnocené. U obecných grafů si již s jednoduchou variantou matice sousednosti nevystačíme. Výhodou je snadná implementace matice sousednosti v PC, nevýhodou matice sousednosti je její velká paměťová náročnost $O(N^2)$. Při práci s relativně malými grafy je velmi mocným nástrojem. S narůstající velikostí grafu se obvykle dostaneme do bodu, kdy nevýhody převáží nad výhodami a musíme použít jinou formu reprezentace grafu.

2.5.2 Seznam hran

Dalším ze způsobů reprezentace grafu v počítači je seznam hran. Jak jsme již zmínili výše, hrana je reprezentována jako uspořádaná nebo neuspořádaná dvojce vrcholů. Přesně tak hrany zaznamenáváme. Zápis seznamu hran jednoduchého grafu na obrázku 2 vypadá následovně:

$$\{(0,1)(1,2)(1,3)(1,4)(2,3)(2,4)(3,4)\}$$

Pokud není síť orientovaná, nezáleží na pořadí vrcholů v zápisu $(0,1) = (1,0)$. Seznamen hran můžeme také zanamenat orientované grafy. Zápis orientováného grafu na obrázku 3 vypadá následovně:

$$\{(0,1)(1,2)(2,3)(3,1)(3,4)(4,1)(4,2)\}$$

Pokud je sít orientovaná, je zapsáno jako první číslo počátečního vrcholu a jako druhé číslo koncového vrcholu. Zde je třeba dbát na pořadí vrcholů $(0,1) \neq (1,0)$! Seznamem hran můžeme reprezentovat i ohodnocené grafy, tím, že k číslům vrcholů se kterými je hrana incidentní, přidáme váhu hrany w_{ij}. Takto zapsaný seznam hran ohodnoceného jednoduchého grafu na obrázku

5 pak vypadá následovně:

$$\{(0,1,3)(1,2,1)(1,3,1)(1,4,2)(2,3,1)(2,4,1)(3,4,1)\}$$

Pokud síť není orientovaná, první dvě čísla představují vrcholy v_i a v_j, se kterými je hrana incidentní, poslední číslo pak představuje váhu hrany w_{ij}. Stejně jako u neorientovaných grafů, můžeme maticí sousedností zaznamenat také ohodnocený orientovaný graf. Zápis ohodnoceného orientovaného grafu na obrázku 6 vypadá následovně:

$$\{(0,1,3)(1,2,1)(2,3,1)(3,1,1)(3,4,1)(4,1,2)(4,2,1)\}$$

Stejně snadno můžeme zaznamenat i obecný graf pomocí seznamu hran. Zápis obecného grafu na obrázku 4 vypadá následovně:

$$\{(0,0)(0,1)(1,0)(1,2)(1,2)(2,1)(2,3)(2,4)(3,1)(3,4)(4,1)(4,2)\}$$

Seznamem hran lze také zaznamenat i ohodnocený či neohodnocený obecný graf. Díky tomu, že seznam hran zaznamenává pouze skutečně existující hrany na rozdíl od matice sousednosti, je to paměťově neúspornější forma reprezentace grafů v PC. Používá se primárně pro uchovávání grafů na perzistentním úložisti.

2.5.3 Seznam sousedů

Dalším způsobem reprezentace grafů, je seznam sousedů. Ten se skládá ze seznamu vrcholů, kde každý vrchol má seznam svých sousedů. Zápis seznamem sousedů jednoduchého grafu na obrázku 2 vypadá následovně:

0	1			
1	0	2	3	4
2	1	3	4	
3	1	2	4	
4	1	2	3	

Na levé straně vidíme seznam vrcholů, na pravé straně má pak každý vrchol seznam sousedních vrcholů.. Můžeme si všimnout, že každá neorientovaná hrana má dva záznamy v seznamu sousedů. Seznamem sousedů můžeme zaznamenat také orientované grafy. Zápis seznamem sousedů orientovaného grafu na obrázku 3 vypadá následovně:

0	1	
1	2	
2	3	
3	1	4
4	1	2

Vrcholy na levé straně reprezentují počáteční vrcholy hran a vrcholy na pravé straně koncové vrcholy hran. Seznamem sousedů můžeme zaznamenávat i ohodnocené grafy. Zápis seznamem sousedů ohodonoceného grafu na obrázku 5 pak vypadá následovně:

0	(1,3)			
1	(0,3)	(2,1)	(3,1)	(4,2)
2	(1,1)	(3,1)	(4,1)	
3	(1,1)	(2,1)	(4,1)	
4	(1,2)	(2,1)	(3,1)	

Jak můžeme vidět, v seznamu vrcholů, se kterými vrchol i sousedí, zapisujeme dvojici údajů. Na prvním místě zapíšeme sousední vrchol j a na druhém místě váhu hrany w_{ij}. Seznamem sousedů můžeme zanamenat také ohodnocené orientované grafy. Zápis seznamem sousedů ohodnoceného orientovaného grafu na obrázku 6 pak vypadá následovně:

0	(1,3)	
1	(2,1)	
2	(3,1)	
3	(1,1)	(4,1)
4	(1,2)	(2,1)

Tak jako seznamem hran i seznamem sousedů můžeme zaznamenat i ohodnocený či neohodnocený obecný graf. Seznam sousedů je nejčastěji používanou formu reprezentace grafu v počítači. Nevýhodou oproti matici sousednosti jsou relativně dlouhé časy pro vyhlednání nebo odebrání hrany, viz. tabulka 1. V programu může být vytvořen v nejjednodušší podobě jako pole polí s proměnnou délkou. Lze však využít sofistikovanějších přístupů, například objektově orientovaný přístup.

2.5.4 Strom sousedů

Newman [4] zmiňuje jako další způsob reprezentace grafů strom sousedů. Tato forma reprezentace má všechny výhody seznamu sousedů, v jistých situacích je však výrazně rychlejší. Strom sousedů využívá podobný koncept jako seznam sousedů. Liší se v tom, že pro zaznamenání sousedních vrcholů nevyužívá seznam, ale strom.

Tabulka 1: Časová složitost operací různých reprezentací grafů s n vrcholy a m hranami. Operace jsou následující: Přidání hrany do sítě (Vložení), odebrání hrany ze sítě (Smazání), otestování zda jsou dané dva vrcholy spojeny hranou (Vyhledání) a získání seznamu sousedů daného vrcholu (Vyčíslení) [4].

Operace	Matice sousednosti	Seznam sousedů	strom sousedů
Vložení	$O(1)$	$O(1)$	$O(\log(m/n))$
Smazání	$O(1)$	$O(m/n)$	$O(\log(m/n))$
Vyhledání	$O(1)$	$O(m/n)$	$O(\log(m/n))$
Vyčíslení	$O(n)$	$O(m/n)$	$O(\log(m/n))$

3 Komplexní sítě a jejich vlasnosti

Věda o sítích je mezioborová disciplína, která kombinuje poznatky z matematiky, fyziky, biologie, informatiky, sociálních věd a mnoha dalších oborů. To vedlo k obohacení vědy o sítích mnoha úhly pohledu vědců z různých disciplín. Zároveň to však ztěžuje situaci, neboť lidské znalosti vědy o sítích jsou rozptýleny napříč vědeckou komunitou. Hlavním zdrojem následující kapitoly je kniha [4] M. E. J. Newmana, která se snaží tyto poznatky sjednotit a prezentovat. Dalším důležitým zdrojem je kniha [1] od A.-L. Barabásiho.

V této kapitole se seznámíme s komplexními sítěmi. Ukážeme si, že mnoho objetků skutečného světa může být považováno za komplexní síť. Povíme si o studovaných vlastnostech komplexních sítí. Seznámíme se se čtyřmi základními kategoriemi sítí - technologickými sítěmi, sociálními sítěmi, informačními sítěmi a biologickými sítěmi. U každé kategorie sítí si představíme ty nejdůležitější zástupce. Na obrázku 8 můžeme vidět ukázku grafické reprezentace komplexní sítě.

Obrázek 8: Ukázka komplexní sítě [11]

3.1 Proč studujeme komplexní sítě?

Mnoho z vědecky zkoumaných systémů se skládá z jednotlivých částí, které jsou nějakým způsobem propojeny. Například internet, který se skládá z počítačů a fyzických propojení mezi nimi. Dalším příkladem může být lidská společnost, která se skládá z lidí a jejich vzájemných vztahů. Mnoho aspektů takových systému stojí za to studovat. Někteří lidé studují povahu jednotlivých komponent systému, například jak fungují počítače nebo jak se lidské bytosti chovají a cítí, zatím co jiní studují povahu propojení a interakcí jednotlivých komponent systému, například komunikační protokoly internetu nebo různé formy mezilidských vztahů. Je zde však ještě třetí aspekt těchto systémů, který je často opomíjený, ale obvykle stěžejní pro pochopení chování daného systému. Tímto aspektem jsou vzory propojení mezi jednotlivými komponentami systému.

3.2 Síť nebo graf?

Vzhledem k tomu, že se ve vědecké literatuře pojmy grafu a sítě používají současně, je třeba si říct, že z hlediska síťové vědy můžeme tyto dva pojmy chápat ekvivalentně. Grafy a sítě používají lehce rozdílné názvosloví. Abychom usnadnili čtenáři orientaci ve vědecké literatuře, přikládáme následujicí tabulku 2, ve které nalezne přehled základního názvosloví, se kterým se může setkat v české a anglické odborné literatuře. Pojmy na jednotlivých řádcích můžeme chápat ekvivaleně.

Tabulka 2: Přehled základních názvosloví [1]

Věda o sítích	Teorie grafů	Network science	Graph teory
Síť	Graf	Network	Graph
Uzel	Vrchol	Node	Vertex
Vazba	Hrana	Link	Edge

A přesto, je zde jistý rozdíl mezi těmito dvěma názvoslovími. Názvosloví Síť-Uzel-Vazba (Network-Node-Link) využíváme když mluvíme o reálných systémech: WWW je síť webových stránek provázaných odkazy, společnost je síť jedinců s rodinnými, přátelskými nebo pracovními vazbami, metabolická síť je soubor všech řetězců chemických reakcí, které probíhají v buňkách. Názvosloví Graf-Vrchol-Hrana (Graph-Vertex-Edge) používáme když mluvíme o matematické reprezentaci těchto sítí: webový graf, sociální graf, metabolický graf. Ne vždy však autoři toto rozdělení názvosloví dodržují a tak jsou často používána jako synonyma.

3.3 Vlastnosti komplexních síti

Jak jsme zmínili, mnoho skutečných systémů může být převedeno na sítě. Pokud se nám podaří získat informace o struktuře těchto sítí, co můžeme následně s těmito daty udělat? Co nám mohou říct o sítích, které zkoumáme? Jaké vlastností sítí jsme schopní změřit a jak souvisejí s realnými úlohami? Jako první příklad si můžeme uvést centralitu. Centralita vcholu se měří různými způsoby, nejjednoduším je měření pomocí stupně vrcholu. Stupeň vrcholu vyjadřuje jeho

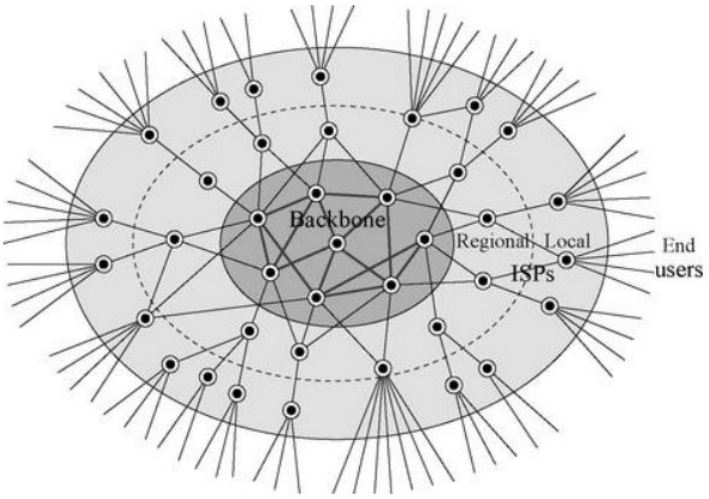
důležitost z hlediska síťové struktury. Čím více má vrchol incidentních hran, tím je důležitější. Studium centrality vrcholů intenzivně využívají sociální vědy. Druhým příkladem může být objev fenoménu "malého světa". Vzdálenost dvěma vrcholy v síti je dána délkou nejkratšího sledu hran mezi těmito vrcholy. Pokud chceme nálezt vzdálenost mezi jakýmikoli dvěma vrcholy v síti, bylo matematicky prokázáno, že průměrná vzdálenost mezi vrcholy v sítí bývá zpravidla velmi malá. Průměrná vzdálenost má obvykle logaritmickou závislost na velikosti sítě (počtu vrcholů). Poprvé byl tento efekt pozorován v sociologické studii [12], kdy vědci přišli se zjištěním, že libovolná dvojice obyvatel USA k sobě může najít cestu prostřednictvím průměrně šesti osob. Tento efekt byl pozorován i na mnoha jiných sítích a pomohl lidem porozumět jak tyto typy sítí fungují. Třetím příkladem je vytváření shluků v sítích. Je obecně známo, že například v sociálních sítích se tvoří komunity (shluky) lidí se stejnými zájmy. Tyto shluky jsou hustě propojené skupiny vrcholů ve větší, řidší síti. Totéž můžeme pozorovat například v obchodních sítích, kde firmy spadající do stejného oboru tvoří shluky. Za předpokladu, že tyto shluky odpovídají nějaké formě společného zájmu, zaměření či jiné vlastnosti, můžeme analýzou sítě zjistit, jaké skupiny se v ní vyskytují. To do jakých skupin jsou sítě rozděleny, nám pomáhá pochopit jak přesně tyto sítě fungují. Studium komunitní struktury sítí je velice aktivní oblastí vědeckého výzkumu.

3.4 Technologické sítě

V následujících řádcích se stručně seznámíme s nejdůležitějšími zástupci technologických sítí.

3.4.1 Internet

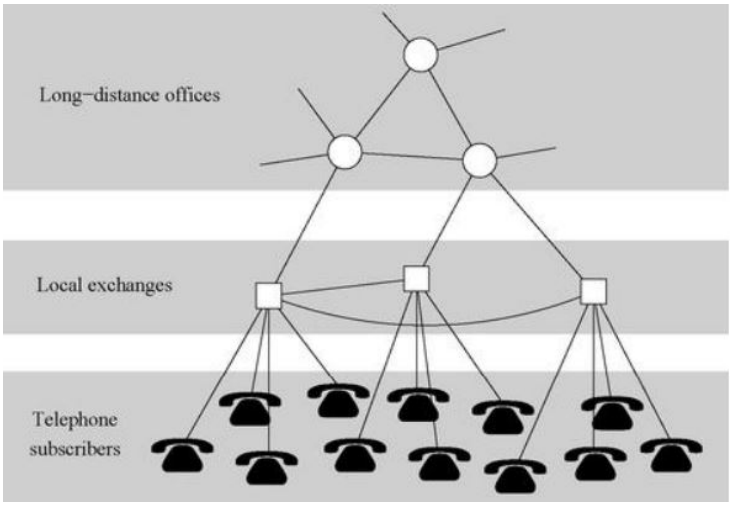
Nejznámější technologickou sítí je bezesporu Internet. Jedná se o celosvětovou síť propojení mezi počítači, routery a ostatními zařízeními. Nejjednoduší síťová reprezentace internetu je taková, kdy vrcholy představují počítače, routery a ostatní zřízení. Hrany pak představují fyzické propojení mezi nimi. V internetu počítače představují "vnější" vrcholy sítě, ze kterých se odesílají a přijímají data, ale nefigurují jako prostředníci pro přenos dat mezi jinými vrcholy. "Vnitřní" vrcholy sítě reprezentují routery (směrovače), speciální zařízení vytvořené pro směrování síťové komunikace z počáteční do cílové destinace. Síť internetu je rozdělena do tří úrovní. Na obrázku 9 můžeme vidět, že tyto tři úrovně si můžeme představit jako tři kruhy. Vnitřní kruh je centrem sítě a nazýváme jej páteřní sítí (z anglického "Backbone"). Střední kruh tvoří poskytovatelé internetu (anglicky "ISP"), kteří zprostředkovávají připojení koncových uživatelů k páteřní síti. Střední kruh se dělí do dvou dalších skupin a to na regionální a lokální poskytovatele internetu. Vnější kruh představuje právě koncové uživatele - domácnosti, kanceláře, školy a mnoho dalších.



Obrázek 9: Schematické zobrazení rozdělení internetu do úrovní [4].

3.4.2 Telefonní sít

Další technologickou sítí, se kterou jsme denně v kontaktu je telefonní síť. Telefonní sítí myslíme síť pozemních a bezdrátových spojení přenášející telefonní hovory. Jedná se o jednu z nejstarších dosud používaných technologických sítí. Základní forma telefoních sítí je relativně jednoduchá. Většíná zemí s pevnou telefoní sítí využívá tříúrovňovou architekturu, kterou můžeme vidět na obrázku 10. Koncoví uživatelé telefonní sítě jsou připojeni přímými linkami k místní telefonní ústředně. Místní ústředny jsou připojeny sdílenými linkami k meziměstským (anglicky "Long distance") ústřednám, které jsou pak dále navzájem propojeny mezi sebou. Struktura telefonních sítí je v mnoha ohledech podobná struktuře Internetu, přestože základní principy funkce zmíněných sítí se liší. Věda o sítích se telefonními sítěmi zabývá o poznání méně než Internetem. Na vině je špatná dostupnost kvalitních dat o jejich struktuře. Přesto, že je topologie telefonních síti dobře známá, společnosti, které telefonní sítě vlastní, tato data veřejně nesdílí. V poslední době lze pozorovat trend, kdy se stále více telefonních hovorů uskutečňuje prostředníctvím Internetu, než po klasických telefonních linkách. To může potenciálně vést k tomu, že se v nedaleké budoucnosti telefonní sítě spojí s Internetem v jednu síť [4].



Obrázek 10: Schematické zobrazení telefonní sítě [4].

3.4.3 Elektrická síť

Elektrické síti se věnovalo několik studií v odborné literatuře. Tyto studie se zabývaly mezinárodní vysokonapěťovou sítí pro přenos elektrické energie. Nízkonapěťové lokání rozvodné sítě jsou obvykle vynechány. Vrcholy takové sítě jsou elektrárny a spínací stanice. Hrany sítě pak odpovídají vysokonapěťovému vedení. Není těžké získat topologii takové sítě, neboť nad ní obvykle dohlíží jedna instituce, která má k dispozici kompletní data. Studiem tohoto typu sítě můžeme získat cenné informace. Podobně jako Internet je i elektrická síť celosvětého rozsahu. Rozmístění vrcholů této sítě je pak zajímavé z geografického, ekonomického a sociálního hlediska. Geografická a topoligická analýza elektrické sítě pomáhá porozumět omezením ovlivňujícím její tvar a další růst. Elektrická sít také vykazuje neobvyklé vzorce chování, jako jsou například kaskádová selhání. To vedlo k překvapivému objevu, že počet a velikosti takových výpadků odpovídá mocninému zákonu (anglicky “power law”) [13].

3.4.4 Přepravní sítě

Značná část vědeckých prací se zabývá přepravními sítěmi. Mezi přepravní sítě patří například letecká doprava, silniční síť nebo železniční síť. Struktura těchto sítí je obvykle snadno dostupná, ale příprava dat bývá zpravidla velmi pracná. Letecké sítě mohou být sestaveny z veřejných seznamů letů, silniční a železniční sítě se sestavují z map. Ve většině studovaných sítí představují vrcholy geografickou polohu a hrany cesty mezi jednotlivými polohami. Ohodnocení přepravních sítí také nebývá jednoduchou disciplínou, neboť mohou být ohodnoceny mnoha různými způsoby, například vzdáleností, počtem přepravních prostředků (letadla, auta, vlaky) využívajících cestu v určitém časovém okně, počtem spojení dopravním prostředkem, množstvím přepraveného nákladu či prostým počtem cestujících. Je proto nutné zvolit ohodnocení, které nejlépe odpovídá problému, který chceme analyzovat.

3.5 Sociální sítě

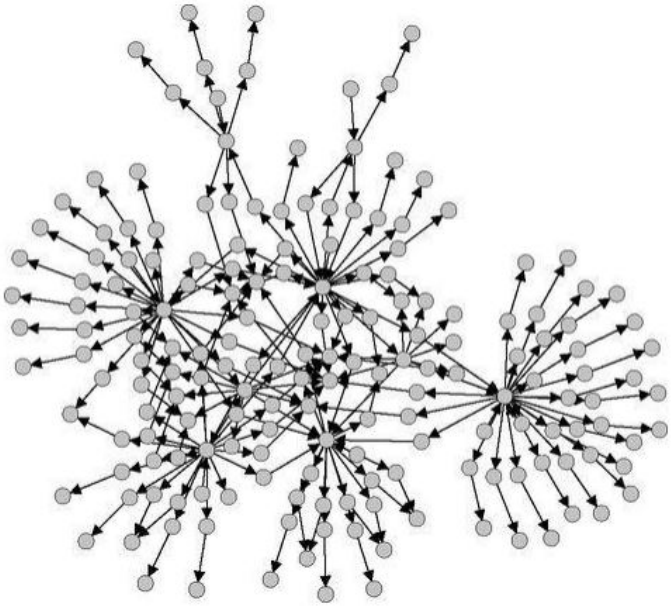
Sociální sítě jsou sítě, ve kterých jsou vrcholy lidé, případně skupiny lidí, a hrany představují nějakou formu sociální interakce mezi nimi, například přátelství. Sociologové označují vrcholy (lidi) jako aktéry a hrany jako vztahy. Většina lidí si pod pojmem sociální síť představí Facebook, Instagram čí jinou online sociální síť. Studium sociálních sítí však sahá mnohem hlouběji do minulosti, kdy tyto online služby neexistovaly. Stejně jako u přepravních sítí i zde mohou být hrany ohodnoceny mnoha různými způsoby. Způsob ohodnocení sociální sítě, který použijeme závisí na tom, jakou otázku chceme zodpovědět. Jako další příklady sociálních sítí si můžeme představit třeba třídu ve škole, sportovní tým, vztahy zaměstnanců ve firmě či rodinu a vztahy mezi jejími členy.

3.6 Informační sítě

Informační sítě představují sítě, složené z bloků dat, které jsou nějakým způsobem provázány. Všechny informační sítě, které známe, jsou vytvořeny lidmi. Bezpochyby nejznámější z nich je World Wide Web (zkráceně WWW). Existují však i jiné informační sítě které stojí za to studovat, například různé citační sítě. Navíc, některé sociální sítě můžeme považovat zároveň za informační sítě, například Facebook, Twitter a podobné. Rozdělení sítí do čtyř základních kategorií tedy není výlučné, některé sítě svým charakterem mohou spadat do více kategorií.

3.6.1 World Wide Web

Přesto, že se v žádném případě nejedná o první informační síť, je pro většinu lidí pravděpodobně tím nejznámějším příkladem informační sítě. Jak jsme již zmínili, vrcholy této sítě jsou webové stránky, které jsou navzájem provázány hypertextovými odkazy. Ty nám umožňují přecházet z jedné webové stránky na jinou. Vzhledem k tomu, že tyto odkazy jsou jednosměrné, jedná se o orientovanou sít. Na obrázku 11 můžeme vidět příklad příklad propojení webových stránek v rámci jednoho webu.



Obrázek 11: Propojení webových stránek v rámci jednoho webu [4]

3.6.2 Citační sítě

Méně známou informační sítí, ovšem mnohem starší, je citační síť vědeckých prací. Mnoho prací, stejně jako tato práce, se odkazuje na jednu či více dřívějších prací jiných autorů v seznamu literatury. Z těchto seznamů v pracích můžeme vytvořit síť, kde jednotlivé práce představují vrcholy a citace odkazující na jiné práce představují orientované hrany. Existuje mnoho důvodů, proč v pracích citovat jiné autory. Jedním z nich je odkázat čtenáře, na další užitečné informace. Dalším důvodem je připsat autorovi citované práce zásluhy. Obecně však platí, že citace indikují, že práce spolu tematicky souvisejí. Citační síť je tedy sítí souvislostí mezi jednotlivými pracemi.

3.6.3 Ostatní informační sítě

Existuje mnoho jiných informačních sítí, avšak žádná z nich se nestala objektem nějaké důkladnější studie. Zmíníme si stručně několik příkladů těchto sítí. Prvním příkladem jsou peer-to-peer sítě, kdy obvykle jedinci přímo sdílí mezi sebou informace, nejčastěji ve formě datových souborů. Často jsou tyto sítě spojovány s nelegálním sdílením autorsky chráněných děl, jako jsou například filmy nebo hudba. Dalším příkladem jsou recenzní sítě, kdy jsou různé produkty hodnoceny na základě uživatelských recenzí. Tyto sítě se v současné době těší čím dál tím větší oblibě. Jako příklad těchto sítí si můžeme uvést asi nejznámější český recenzní server Heureka (`https://www.heureka.cz/`) či Česko-Slovenskou filmovou databázi (`https://www.csfd.cz/`).

3.7 Biologické sítě

Sítě jsou intenzivně využívány v mnoha odvětvích biologie jako vhodná reprezentace vzorů interakcí mezi biologickými prvky. Molekulární biologové používají sítě k reprezentaci vzorů reakcí probíhajících mezi chemickými prvky v buňkách. Neurologové využívají sítě k zachycení propojení mezi buňkami v mozku. Ekologové zase studují vzory interakcí mezi jednotlivými druhy živočichů v ekosystémech.

3.7.1 Biochemické sítě

Mezi biologickými sítěmi se biochemické sítě těší největší pozornosti vědecké komunity. Biochemické sítě reprezentují vzory interakcí a řídících mechanizmů na molekulární úrovni uvnitř buněk. Hlavními zástupci těchto sítí jsou metabolické, proteinové a genetické regulační sítě.

Metabolismus je chemický proces, při němž buňky rozkládají přijaté živiny na jiné užitečné stavební bloky a pak je opět skládají dohromady aby vytvořily biologické molekuly, které buňky potřebují. Tento rozklad a opětovné složení obvykle zahrnuje spoustu dílčích kroků (chemických reakcí), které nakonec přetvoří vstupní látky na výsledné užitečné produkty. Množina všech těchto dílčích reakcí pro různé varianty vstupních látek představuje metabolickou síť. Vrcholy této sítě jsou chemické sloučeniny, které vznikají a zanikají při reakcích. Samotné chemické reakce pak představují hrany této sítě.

3.7.2 Neurologické sítě

Nové využití sítí v biologii přinesla studie mozku a centrální nervové soustavy živočichů. Jednou z primárních funkcí mozku je zpracovávat informace. Primárním elementem v mozku pro zpracování inforací je neuron, specializovaná mozková buňka, která má obvykle několik informačních vstupů a jeden výstup. V závislosti na živočichovi, může mozek obsahovat od několika neuronů až po stovky miliónů. Navzájem propojené neurony tvoří neurologickou síť s obdivuhodnými schopnosti rozhodovat a počítat [4]. Neurony v této síti představují vrcholy, které mohou být propojeny dvěma různými typy orientovaných hran. Existuje mnoho různých typů neuronů, které se v praxi dají reprezentovat různými typy vrcholů. Získávání dat o těchto sítích je velice náročné, neboť neexistují přímé experimentální techniky pro získávání jejich struktury.

3.7.3 Ekologické sítě

Živočichové v ekosystému se mohou navzájem ovlivňovat mnoha různými způsoby. Mohou jíst jeden druhého, mohou parazitovat jeden na druhém, mohou spolu soutěžit o zdroje či mít vzájemně prospěšné vztahy. Přesto, že celá tato síť by se dala zachytit jako jedna rozsáhlá síť s různými typy hran, ekologové ji tradičně rozdělují podle typů interakcí do menších sítí. Nejznámější a nejdéle studovanou z těchto sítí je jistě potravinový řetězec, který zachycuje vztahy mezi predátory a kořistí (kdo koho jí). Jedná se o orientovanou síť, kde živočichové nebo jejich

skupiny reprezentují vrcholy a orientované hrany představují vztah predátor-kořist. Většina lidí by při kreslení grafu takovéto sítě intuitivně nakreslila šípku směrem od predátora ke kořisti, ale ekologové je zakreslují naopak, směrem od kořisti k predátorovi.

4 Shlukovací koeficient

Velké množství sítí vykazuje tendenci vytvářet trojúhelníky (úplné grafy na 3 vrcholech). Topologie těchto sítí se odchyluje od náhodných sítí, ve kterých se trojúhelníky tvoří vzácně. Tato tendence se nazývá shlukování. Počátky shlukování můžeme najít v sociologii, kde se používají podobné koncepty. V typické sociální síti je velká pravděpodobnost, že dva přátelé téhož člověka se také navzájem znají. Shlukování okolo vrcholu i je vyjádřeno shlukovacím koeficientem C_i. V následujících řádcích si představíme základní (neohodnocenou, neváženou) variantu výpočtu shlukovacího koeficientu a poté se seznámíme s váženými (ohodnocenými) variantami výpočtu. Vážené varianty shlukovacího koeficietu budou odlišeny vlnovkou $\tilde{C}_i$. Je nutno zmínit, že u všech variant výpočtů shlukovacího koeficientu se předpokládá, že se jedná o neorientovanou síť bez multihran a smyček. Hlavním zdrojem této kapitoly je článek od J. Saramäkiho [9].

4.1 Nevážený shlukovací koeficient

V neorientovaných neohodnocených sítích je slukovací koeficient C_i vrcholu i definován jako počet trojúhelníků, kterých je vrchol i součástí, děleno maximálním možným počtem takových trojúhelníků:

$$C_i = \frac{t_i}{\binom{k_i}{2}} = \frac{2t_i}{k_i(k_i-1)} \tag{1}$$

kde t_i udává počet trojúhelníků v okolí vrcholu i a k_i je stupeň vrcholu i. V orientovaných sítích je vzorec pro výpočet shlukovacího koeficientu vyjádřen následovně:

$$C_i = \frac{t_i}{k_i(k_i-1)} \tag{2}$$

.

Algoritmus 1 Výpočet neváženého shlukovacího koeficientu vrcholu i

Require: G
  $C_i \leftarrow 0$
  if $k_i > 1$ **then**
    $t_i \leftarrow 0$
    for all $j \in N_i(G)$ **do**
      for all $l \in N_i(G)$ **do**
        if $\{j;l\} \in E(G)$ **then**
          $t_i \leftarrow t_i + 1$
        end if
      end for
    end for
    $C_i \leftarrow t_i/(k_i * (k_i - 1))$
  end if

4.2 Shlukovací koeficient podle Barrata

Barrat [5] byl první, kdo ve své práci navrhl váženou variantu shlukovacího koeficientu. Ve svém výpočtu využívá vážený stupeň vrcholu s_i[6], který definuje jako součet vah všech hran, se kterými je vrchol i incidentní. To lze zapsat následovně: $s_i = \sum_{j \in V(N_i)} w_{ij}$, kde w_{ij} představuje váhu hrany mezi vrcholy i a j. Vážený shlukovací koeficient podle Barrata budeme značit $\tilde{C}_{i,B}$ a definice jeho výpočtu je následující:

$$\tilde{C}_{i,B} = \frac{1}{s_i(k_i - 1)} \sum_{j,l \in V(N_i)} \frac{w_{ij} + w_{il}}{2} a_{ij} a_{jl} a_{il} \tag{3}$$

kde $a_{ij} = 1$ pokud existuje hrana mezi vrcholy i a j, jinak $a_{ij} = 0$. Za předpokladu, že $s_i = k_i(s_i/k_i) = k_i\langle w_i \rangle$ můžeme tento vzorec přepsat následovně:

$$\tilde{C}_{i,B} = \frac{1}{k_i(k_i - 1)} \sum_{j,l \in V(N_i)} \frac{1}{\langle w_i \rangle} \frac{w_{ij} + w_{il}}{2} a_{ij} a_{jl} a_{il} \tag{4}$$

kde $\langle w_i \rangle = \sum_j w_{ij}/k_i$. Tato přepsaná forma jasně ukazuje, že přispění každého trojúhelníku závisí na poměru průměru dvou sousedních hran vůči průměrné váze hrany.

Algoritmus 2 Výpočet váženého shlukovacího koeficientu vrcholu i podle Barrata

```
Require: G
  C̃_{i,B} ← 0
  if k_i > 1 then
    sum ← 0
    for all j ∈ N_i(G) do
      for all l ∈ N_i(G) do
        if {j; l} ∈ E(G) then
          sum ← sum + (w_ij + w_il)/2
        end if
      end for
    end for
    C̃_{i,B} ← sum/(s_i * (k_i − 1))
  end if
```

4.3 Shlukovací koeficient podle Onnely

Onnela [6] ve své práci navrhl váženou variantu shlukovacího koeficentu založenou na konceptu "podgrafové intenzity", definované jako geometrický průměr ohodnocení hran v podgrafu. Vážený shlukovací koeficient podle Onnely budeme značit $\tilde{C}_{i,O}$ a definice jeho výpočtu je následující:

[6] Z anglického strenght (síla)

$$\tilde{C}_{i,O} = \frac{1}{k_i(k_i-1)} \sum_{j,l \in V(N_i)} (\hat{w}_{ij}\hat{w}_{il}\hat{w}_{jl})^{1/3} \tag{5}$$

Zde jsou všechny váhy hran normalizovány nejvyšší váhou hrany v síti, $\hat{w}_{ij} = w_{ij}/max(w)$ a přispění každého trojúhelníku závísí na vahách všech tří hran. Přispění trojúhelníků u nichž je váha jedné hrany zanedbatelná, je pak také zanedbatelné.

Algoritmus 3 Výpočet váženého shlukovacího koeficientu vrcholu i podle Onnely

Require: $G; max(w)$
  $\tilde{C}_{i,O} \leftarrow 0$
  if $k_i > 1$ **then**
    $sum \leftarrow 0$
    for all $j \in N_i(G)$ **do**
      for all $l \in N_i(G)$ **do**
        if $\{j; l\} \in E(G)$ **then**
          $sum \leftarrow sum + \sqrt[3]{w_{ij}w_{il}w_{jl}}/max(w)$
        end if
      end for
    end for
    $\tilde{C}_{i,O} \leftarrow sum/(k_i * (k_i - 1))$
  end if

4.4 Shlukovací koeficient podle Zhanga

Zhang [7] definoval ve své práci vážený shlukovací koeficient, vzhledem ke genovým sítím. Vážený shlukovací koeficient podle Zhanga budeme značit $\tilde{C}_{i,Z}$ a definice jeho výpočtu je následující:

$$\tilde{C}_{i,Z} = \frac{\sum_{j,l \in V(N_i)} \hat{w}_{ij}\hat{w}_{il}\hat{w}_{jl}}{(\sum_l \hat{w}_{il})^2 - \sum_l \hat{w}_{il}^2} \tag{6}$$

Stejně jako výše, i zde jsou váhy všech hran normalizovány maximalní váhou hrany v síti $max(w)$. Tuto rovnici můžeme také upravit následovně:

$$\tilde{C}_{i,Z} = \frac{\sum_{j,l \in V(N_i)} \hat{w}_{ij}\hat{w}_{il}\hat{w}_{jl}}{\sum_{j \neq l} \hat{w}_{ij}\hat{w}_{il}} \tag{7}$$

Algoritmus 4 Výpočet váženého shlukovacího koeficientu vrcholu i podle Zhanga

Require: $G; max(w)$
  $\tilde{C}_{i,Z} \leftarrow 0$
  if $k_i > 1$ **then**
    $sum1 \leftarrow 0$
    $sum2 \leftarrow 0$
    for all $j \in N_i(G)$ **do**
      for all $l \in N_i(G)$ **do**
        $temp \leftarrow w_{ij}w_{il}/max(w)^2$
        if $\{j; l\} \in E(G)$ **then**
          $sum1 \leftarrow sum1 + temp * w_{jl}/max(w)$
        end if
        if $j \neq l$ **then**
          $sum2 \leftarrow sum2 + temp$
        end if
      end for
    end for
    $\tilde{C}_{i,Z} \leftarrow sum1/sum2$
  end if

4.5 Shlukovací koeficient podle Holmeho

Holme [8] definoval ve své práci vážený shlukovací koeficient velice podobným způsobem. Vážený shlukovací koeficient podle Holmeho budeme značit $\tilde{C}_{i,H}$ a definice jeho výpočtu je následující:

$$\tilde{C}_{i,H} = \frac{\sum_{j,l \in V(N_i)} w_{ij}w_{il}w_{jl}}{max(w)\sum_{j,l} w_{ij}w_{li}} \tag{8}$$

V jeho rovnici však odpadá nutnost podmínky $j \neq l$ u sumy ve jmenovateli.

Algoritmus 5 Výpočet váženého shlukovacího koeficientu vrcholu i podle Holmeho

Require: $G; max(w)$
  $\tilde{C}_{i,H} \leftarrow 0$
  if $k_i > 1$ **then**
    $sum1 \leftarrow 0$
    $sum2 \leftarrow 0$
    for all $j \in N_i(G)$ **do**
      for all $l \in N_i(G)$ **do**
        $temp \leftarrow w_{ij}w_{il}$
        if $\{j; l\} \in E(G)$ **then**
          $sum1 \leftarrow sum1 + temp * w_{jl}$
        end if
        $sum2 \leftarrow sum2 + temp$
      end for
    end for
    $\tilde{C}_{i,H} \leftarrow sum1/(max(w) * sum2)$
  end if

4.6 Porovnání vážených variant shlukovacího koeficientu

Tabulka 3: Porovnání motivace a vybraných vlastností různých shlukovacích koeficientů [9]

Koeficient	Motivace
$\tilde{C}_B$	Odraží jaká část vah hran je součátní okolních trojúhelníků.
$\tilde{C}_O$	Odráží jak velké jsou váhy trojúhelníků v porovnání se síťovým maximem.
$\tilde{C}_Z$	Založený pouze na vahách hran, necitlivý na přídavný šum v podobě "falešně pozitivních hran" s malými vahami.
$\tilde{C}_H$	Podobný jako $\tilde{C}_Z$, založený pouze na vahách hran.

Vlastnost	$\tilde{C}_B$	$\tilde{C}_O$	$\tilde{C}_Z$	$\tilde{C}_H$
1. $\tilde{C} = C$ pokud přiřadíme všem vahám hran stejnou hodnotu	X	X	X	
2. $\tilde{C} \in [0, 1]$	X	X	X	
3. Používá nejvyšší váhu hrany v síti $max(w)$ při normalizaci		X	X	X
4. Zohledňuje váhy všech hran v trojúhelnících jejichž součástí je i		X		X
5. Nezávislý na uspořádání vah hran v trojúhelníku		X		
6. Zohledňuje váhy hran, které nejsou součástí žádného trojúhelníku	X		X	X

Tabulka 3 prezentuje vybrané vlastnosti a rozdíly čtyř zmíněných variant váženého shlukovacího koeficientu. Tyto vlastnosti si dále blíže rozebereme.

1. $\tilde{C} = C$ pokud přiřadíme všem vahám hran stejnou hodnotu. Tato podmínka je splněna všemi koeficienty kromě $\tilde{C}_H$. Když všem hranám přiřadíme stejnou hodnotu, pak $\tilde{C}_H = t_i/k_i^2$, což se blíží C pouze když $k \gg 1$.

2. $\tilde{C} \in [0, 1]$. To platí pro všechny koeficienty kromě $\tilde{C}_H$, který nikdy nedosáhne 1 ze stejného důvodu jako v předchozím bodě. Rozeberme limitní hodnoty ($\tilde{C} = 0, \tilde{C} = 1$) detailněji. Pro všechny vážené shlukovací koeficienty platí, že $\tilde{C} = 0$ vyjadřuje absenci trojúhelníků. Nutnou podmínkou pro $\tilde{C}_B = 1$, $\tilde{C}_O = 1$ a $\tilde{C}_Z = 1$ je existence hran mezi všemi sousedními vrcholy vrcholu i. Nicméně každý vážený shlukovací koeficient má jiné nároky na váhy hran. Pokud $C_i = 1$ pak i $\tilde{C}_B = 1$ nezávisle na vahách hran. Naopak $\tilde{C}_O = 1$ vyžaduje, aby se váhy všech hran v trojúhelnících rovnaly globálnímu maximu $max(w)$. Nakonec $\tilde{C}_Z = 1$ pokud každá "vnější" hrana w_{jl} trojuhelníku je rovna globálnímu maximu $max(w)$.

3. Používá nejvyšší váhu hrany v síti $max(w)$ při normalizaci. To je pravda pro všechny varianty vážených shlukovacích koeficientů kromě $\tilde{C}_B$, kde se bere v potaz vážený stupeň vrcholu s_i. Tato konkrétní volba znamená, že ve stejné síti dva různé vrcholy s podobnou topologií sousedů a podobným poměrem vah hran mohou mít podobný $\tilde{C}_B$, i přesto že hrany v okolí jednoho vrcholu mají relativně malou váhu, zatímco hrany v okolí druhého vrcholu jsou výrazně větší.

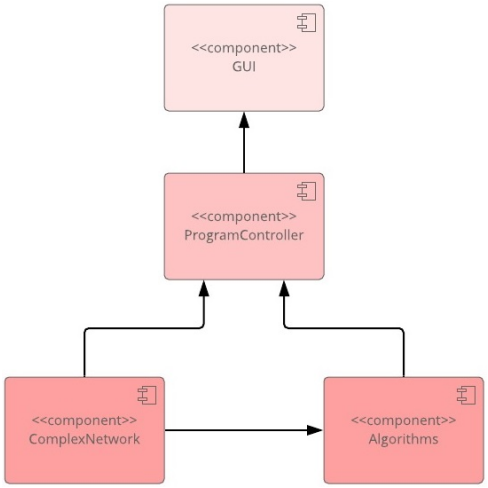
4. Zohledňuje váhy všech hran v trojúhelnících jejichž součástí je i. To platí pro $\tilde{C}_O$ a $\tilde{C}_H$. Nicméně $\tilde{C}_B$ zohledňuje pouze váhy hran incidentních s i.

5. Nezávislý na uspořádání vah hran v trojúhelníku. Tuto vlastnost má pouze $\tilde{C}_O$.

6. .Zohledňuje váhy hran, které nejsou součástí žádného trojúhelníku. To platí pro všechny koeficienty kromě $\tilde{C}_O$, kde tyto hrany vstupují do výpočtu jako součást k_i.

5 Implementace

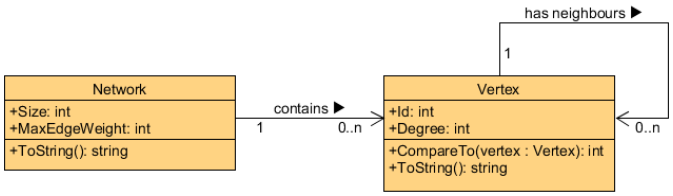
Aplikace je napsána jazyce C# s využitím technologie .Net. Aplikace se skládá ze čtyř komponent (ComplexNetwork, Algorithms, ProgramController, GUI). Schematické propojení komponent můžeme vidět na obrázku 12. Jádro tvoří komponenta ComplexNetwork, která poskytuje datovou strukturu pro reprezentaci sítě, a komponenta Algorithms, která postkytuje výpočty shlukovacího koeficientu. Komponenta ProgramControler zpracovává vstupy od uživatele a poskytuje vypočtená data. Komponenta GUI poskytuje uživateli intuitivní rozhraní pro načítání, ukládání a vizualizaci dat 15. Vizualice dat je realizována pomocí knihoven OxyPlot (`https://www.oxyplot.org/`).



Obrázek 12: Diagram komponent

5.1 Datová struktura pro reprezentaci sítě

S ohledem na fakt, že časově nejnáročnější operací při výpočtech shlukovacího koeficientu je vyhledávání hran, byla použita struktura strom sousedů (2.5.4) s objektově orientovaným přístupem. Implementace této datové struktury je v komponentě ComplexNetwork. Na obrázku 13 můžeme vidět diagram tříd této datové struktury.



Obrázek 13: Diagram tříd reprezentujících síť

Třída Vertex představuje vrchol sítě. Ve výpisu 1 si můžeme všimnout, že pro záznam sousedů používá SortedDictionary, který je v jazyku C# implementován jako binární strom.

```
public class Vertex : IComparable<Vertex>
{
   public int Id { get; }
   public SortedDictionary<Vertex, double> Neighbours { get; }
   public int Degree
   {
      get { return this.Neighbours.Count; }
   }

   public Vertex(int id, SortedDictionary<Vertex, double> neighbours)
   {
      this.Id = id;
      this.Neighbours = neighbours;
   }
   public int CompareTo(Vertex vertex)
   {
      return this.Id - vertex.Id;
   }
}
```

Výpis 1: Třída Vertex

Ve výpisu 2 můžeme vidět třídu Network, která reprezentuje síť jako celek. Obě dvě třídy obsahují pouze informace nutné pro výpočty.

```
public class Network
{
   public int Size { get; internal set; }
   public SortedDictionary<int, Vertex> Vertices { get; }
   public double MaxEdgeWeight { get; internal set; }

   public Network()
   {
      this.Size = 0;
      this.Vertices = new SortedDictionary<int, Vertex>();
   }
}
```

Výpis 2: Třída Network

Komponenta Complex network obsahuje také statickou třídu Loader, která slouží ke korektnímu načtění zdrojových souborů sítí z perzistentního úložiště do paměti programu ve formě třídy Network. Podporovány jsou formáty CSV, TNET a UCINET.

5.2 Výpočty shlukovacího koeficientu

Komponenta Algorithms obsahuje statickou třídu ClusteringCoefficient, ve které jsou implementovány výpočty shlukovacího koeficientu. Základ tvoří 5 metod: CalculateUnweighted (výpis 3), CalculateBarrat (výpis 4), CalculateOnnela (výpis 5), CalculateZhang (výpis 6) a CalculateHolme (výpis 7). Tyto metody přijímají jako argumet třídu network a vracejí seznam vrcholů a jejich shlukovacích koeficientů. Zavoláním metody Distribution (výpis 8), která přijímá jako argument seznam vrcholů a jejich shlukovacích koeficientů a třídu network pak můžeme získat distribuci průměrného shlukovacího koeficientu vůči stupni vrcholu. Pokud se podíváme na jednotlivé výpisy kódů výpočtů (3, 4, 5, 6 a 7), můžeme si všimnout, že jsou velmi podobné a při všech výpočtech probíhá stejná vyhledávací sekvence. Pro potřeby hromadného výpočtu všech typů koeficientů proto byla vytvořena metoda CalculateAll, která tohoto faktu využívá a počítá všech pět typů shlukovacího koeficientu včetne distribuce průměrného shlukovacího koeficientu vůči stupni vrcholu zároveň při jednom průchodu. Výpočet touto optimalizovanou metodou je přibližně čtyřikrát rychlejší, než kdybychom postupně volali metody CalculateUnweighted, CalculateBarrat, CalculateOnnela, CalculateZhang a CalculateHolme.

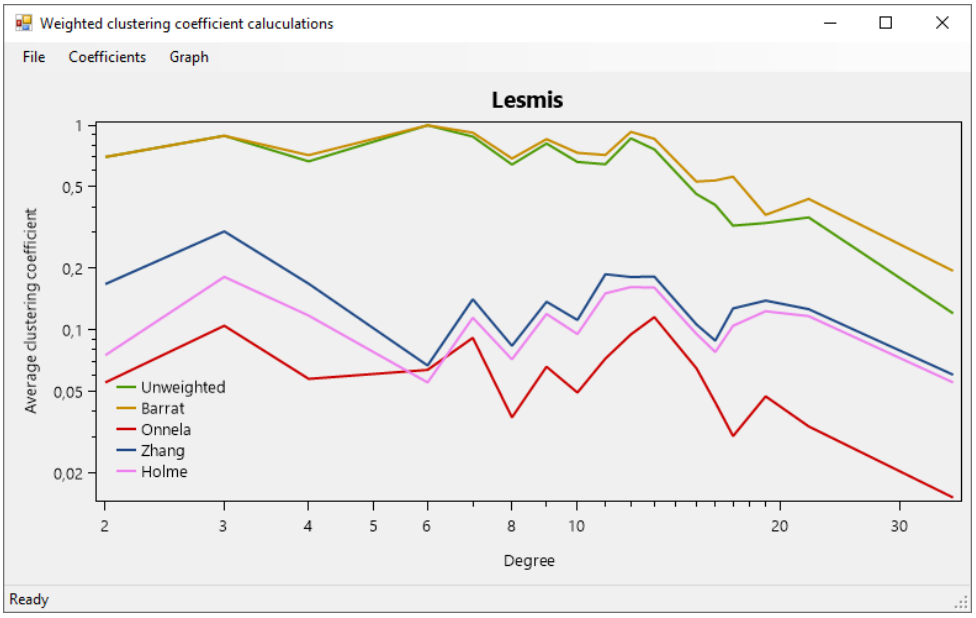
5.3 Uživatelské rozhraní

Komponenty ProgramController a GUI poskytují uživateli jednoduché rozhraní pro načítání dat, ukladání výsledků a jejich zobrazení. Náhled tohoto rozhraní můžeme vidět na obrázku 15. Rozhraní je vytvořeno s využitím knihoven Windows Forms a grafy pro vizualizaci dat jsou kresleny pomocí knihovny OxyPlot. Na obrázku 14 můžeme vidět případy užití.

Obrázek 14: Use case diagram

Jednotlivé případy užití si dále rozebereme:

- Načtení sítě probíhá formou dialogového okna, kde si uživatel zvolí typ souboru a poté samotný soubor se sítí.
- Uložení koeficientů probíhá opět formou dialogového okna, ve kterém si uživatel zvolí název CSV souboru, do kterého budou koeficienty uloženy.
- Uložení distribuce průměrného shlukovacího koeficientu vůči stupni vrcholu probíhá taktéž formou dialogového okna, kde si uživatel zvolí typ souboru (CSV nebo PNG), do kterého bude distribcue uložena, a pak jeho název. Pokud si uživatel zvolí soubor ve formátu PNG, graf je uložen ve stejné velikosti, jak jej vidí uživatel v aplikaci.
- Uživatel si v nabídce koeficientů může zaškrtáváním zvolit, které typy koeficientů budou zobrazeny a ukládány.
- Uživatel si může v nabídce zvolit, zda budou osy grafu zobrazeny v lineárním nebo logaritmickém měřítku.
- Uživatel si může v nabídce zvolit, ve kterém rohu grafu bude zobrazena legenda.



Obrázek 15: Uživatelské rozhraní

6 Experimenty

6.1 Ověření správnosti implementace

Prvním provedeným experimentem bylo ověření správnosti implementace. Jako vzor posloužil ohodnocený jednoduchý graf na obrázku 5. Výsledky všech variant výpočtu byly nejdříve ručně ověřeny. Poté byl graf načten přes uživatelské rozhraní a výsledky byly spočítány aplikací. Vyšly totožně.

Tabulka 4: Vypočtené shlukovací koeficienty ohodnoceného jednoduchého grafu na obrázku 5

i	C_i	$\tilde{C}_{i,B}$	$\tilde{C}_{i,O}$	$\tilde{C}_{i,Z}$	$\tilde{C}_{i,H}$
0	0.00000	0.00000	0.00000	0.00000	0.00000
1	0.50000	0.38095	0.19555	0.09804	0.06803
2	1.00000	1.00000	0.36221	0.44444	0.29630
3	1.00000	1.00000	0.36221	0.44444	0.29630
4	1.00000	1.00000	0.39109	0.33333	0.20833

Poté proběhl další krok ověření, kde jako vzor posloužil jednoduchý graf na obrázku 2. Na základě tvrzení v tabulce 3 jsme očekávali, že pokud jsou výpočty správně implementovány, vyjou $\tilde{C}_{i,B}$, $\tilde{C}_{i,O}$ a $\tilde{C}_{i,Z}$ pro neohodnocený graf stejně jako C_i. Výsledky spočítané aplikací toto očekávání splnily.

Tabulka 5: Vypočtené shlukovací koeficienty jednoduchého grafu na obrázku 2

i	C_i	$\tilde{C}_{i,B}$	$\tilde{C}_{i,O}$	$\tilde{C}_{i,Z}$	$\tilde{C}_{i,H}$
0	0.00000	0.00000	0.00000	0.00000	0.00000
1	0.50000	0.50000	0.50000	0.50000	0.37500
2	1.00000	1.00000	1.00000	1.00000	0.66667
3	1.00000	1.00000	1.00000	1.00000	0.66667
4	1.00000	1.00000	1.00000	1.00000	0.66667

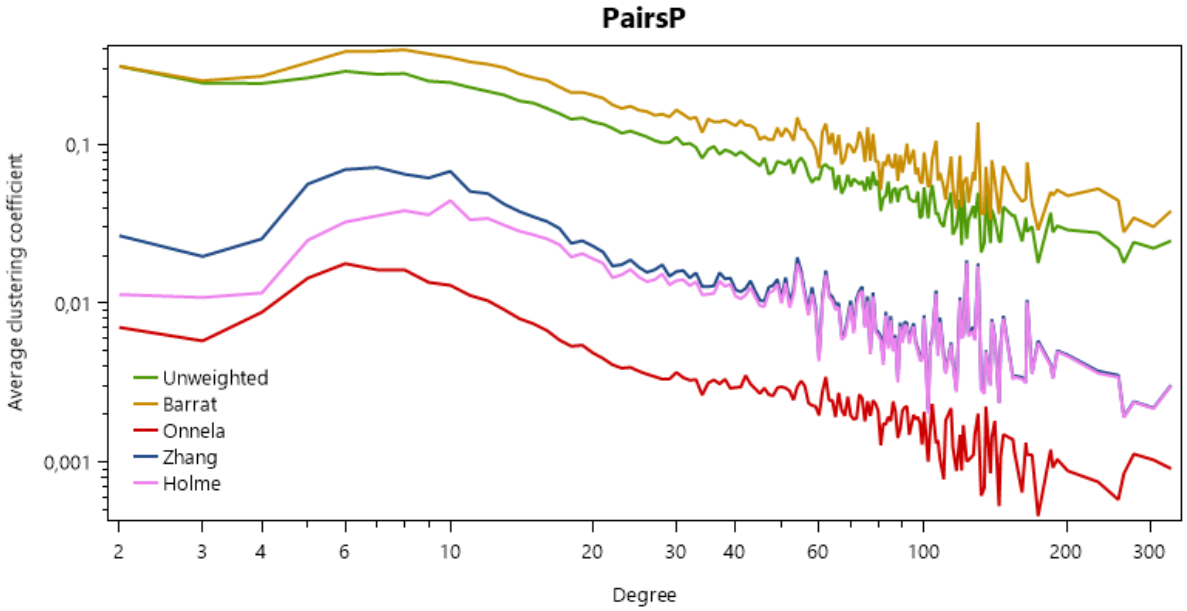
6.2 Zobrazení vztahu mezi průměrným shlukovacím koeficientem a stupněm vrcholu

Dalším provedeným experimentem bylo zobrazení vztahu mezi průměrným shlukovacím koeficientem a stupněm vrcholu. Prostřednictvím vytvořené aplikace byly níže zmíněné datové sady zpracovány. Následuje seznam použitých datových sad s výsledky.

6.2.1 Síť volných asociací slov

Síť volných asociací slov (zkráceně označována PairsP) je informační síť, která byla vytvořena na základě odpovědí více než 6000 respondentů. Respondendům byla předkládána výbraná slova

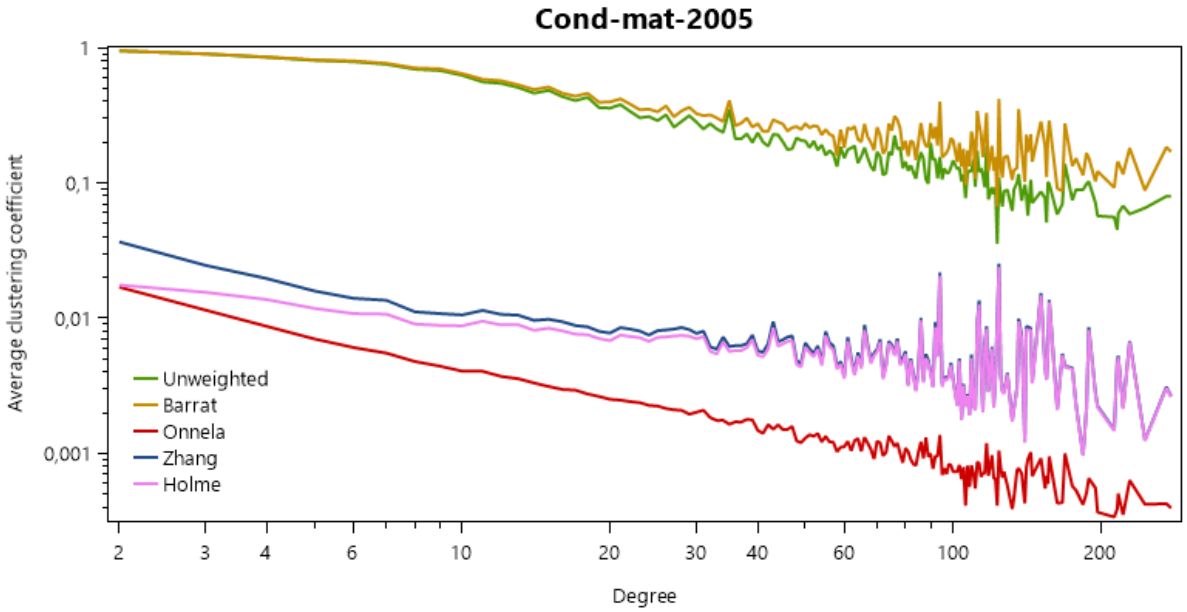
a jejich úkolem bylo napsat první slovo, které je napadne v souvislosti s předloženým slovem. Více o tom, jak byla tato síť zpracována můžeme najít v článku [20]. Vrcholy této sítě představují jednotlivá slova a hrany reprezentují asociaci mezi nimi. Ohodnocení hran představuje počet stejných asociací na základě odpovědí od respondentů. Z odpovědí respondentů byla vytvořena orientovaná síť asociací. Pro účely našeho experimentu, byla orientace hran zanedbána. Duplicitní hrany byly sečteny a smyčky vynechány. Po našem předzpracování má výsledná neorientovaná síť 10617 vrcholů a 63786 hran.



Obrázek 16: Závislost průměrného shlukovacího koeficientu na stupni vrcholu (PairsP)

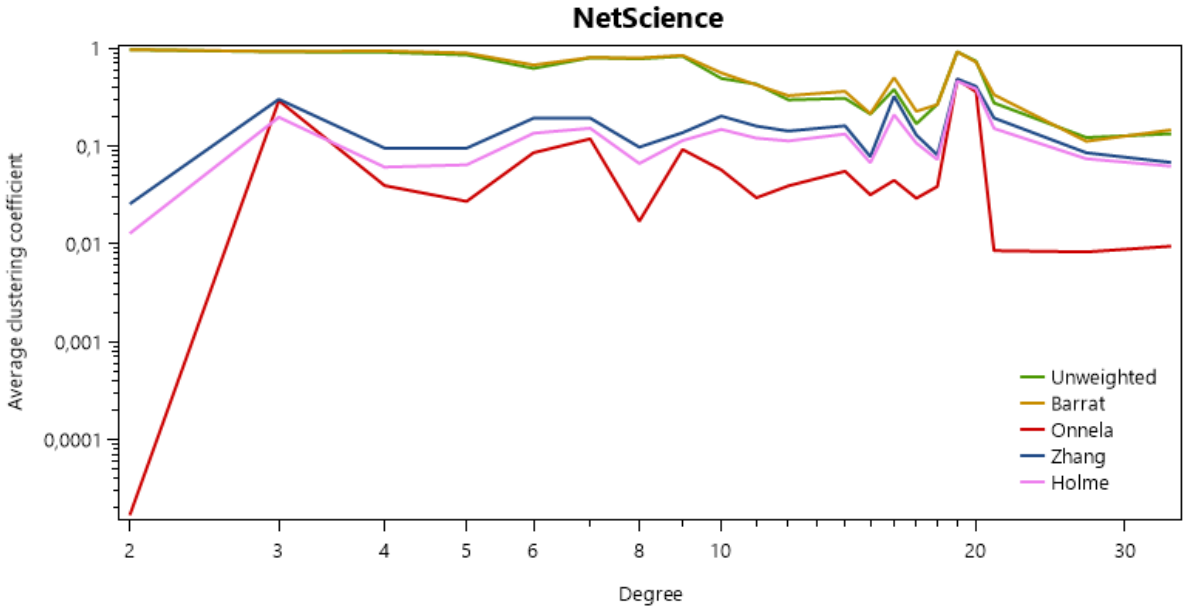
6.2.2 Sítě vědecké spolupráce

Budeme pracovat se třemi datasety reprezentujícími síť vědecké spolupráce. Všechny tři sítě jsou sociálními a zároveň informačními sítěmi. První dataset (zkráceně označovaný cond-mat-2005) je síť mapující spolupráci mezi vědci sdílejícími předtisky svých prací v archivu `https://arxiv.org/archive/cond-mat` mezi daty 1.1.1995 až 31.3.2005. Jedná se o neorientovanou síť s 40420 vrcholy a 175692 hranami. Dataset byl zpracován M. E. J. Newmanem a postupy použité při jeho zpracování jsou posány v článku [14].



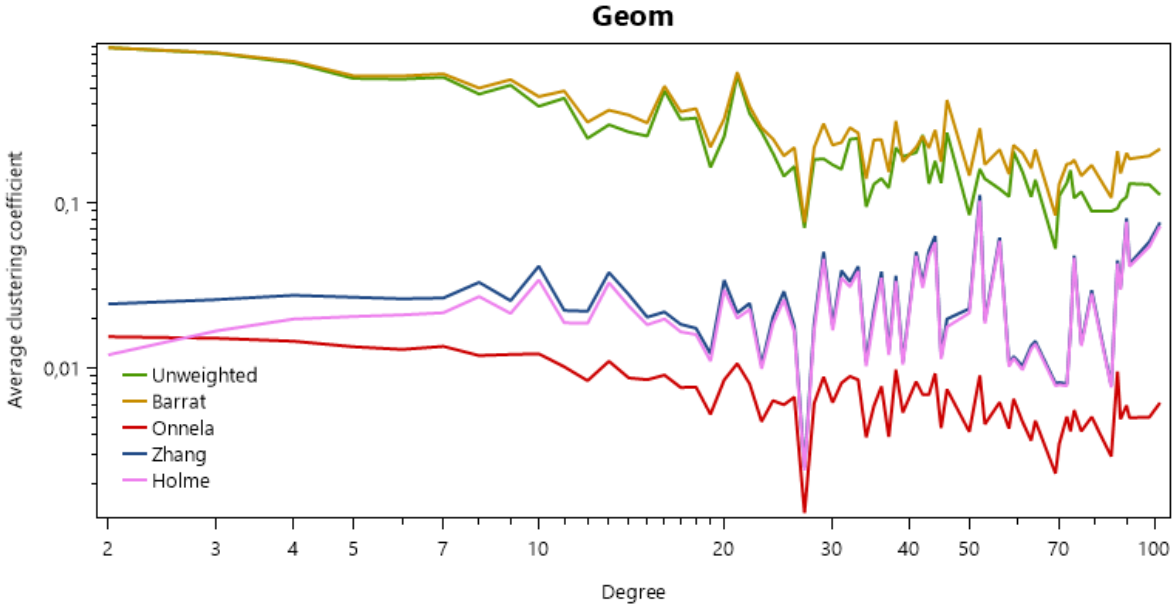
Obrázek 17: Závislost průměrného shlukovacího koeficientu na stupni vrcholu (cond-mat-2005)

Druhý dataset (zkráceně označovaný NetScience) je síť mapující spolupráci vědců v oblasti vědy o sítích. Jedná se o neorientovanou síť s 1589 vrcholy a 2742 hranami. Dataset byl zpracován M. E. J. Newmanem v roce 2006 a použit v jeho článku [19].



Obrázek 18: Závislost průměrného shlukovacího koeficientu na stupni vrcholu (NetScience)

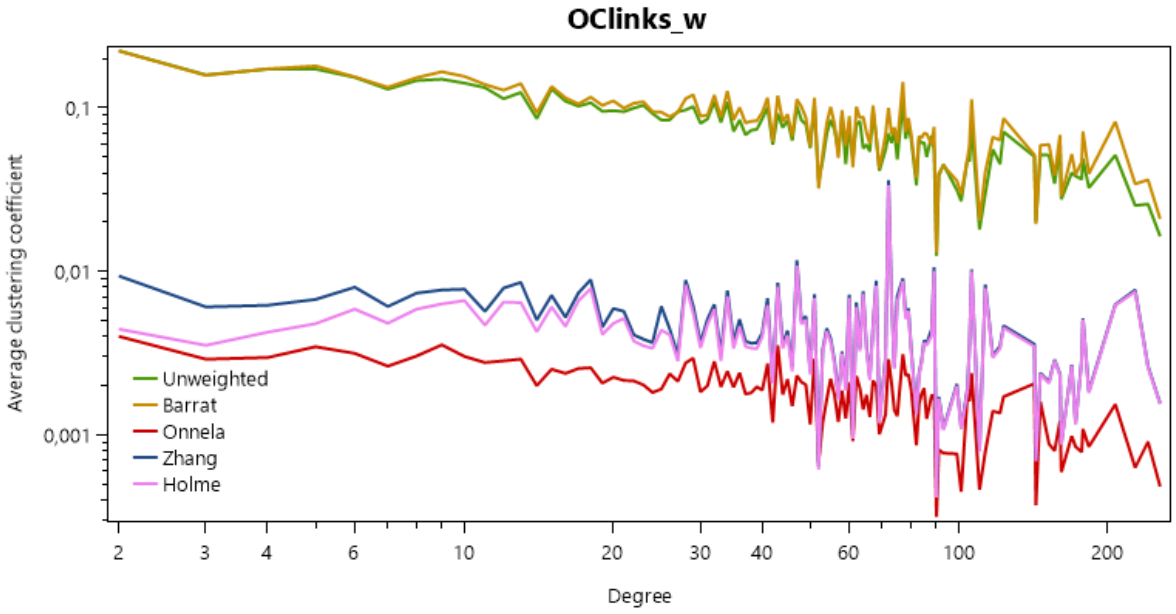
Třetí dataset (zkráceně označovaný Geom) je síť mapující spolupráci vědců v oblasti výpočetní geometrie. Jedná se o neorientovanou síť s 7343 vrcholy a 11898 hranami.



Obrázek 19: Závislost průměrného shlukovacího koeficientu na stupni vrcholu (Geom)

6.2.3 Sociální síť podobná Facebooku

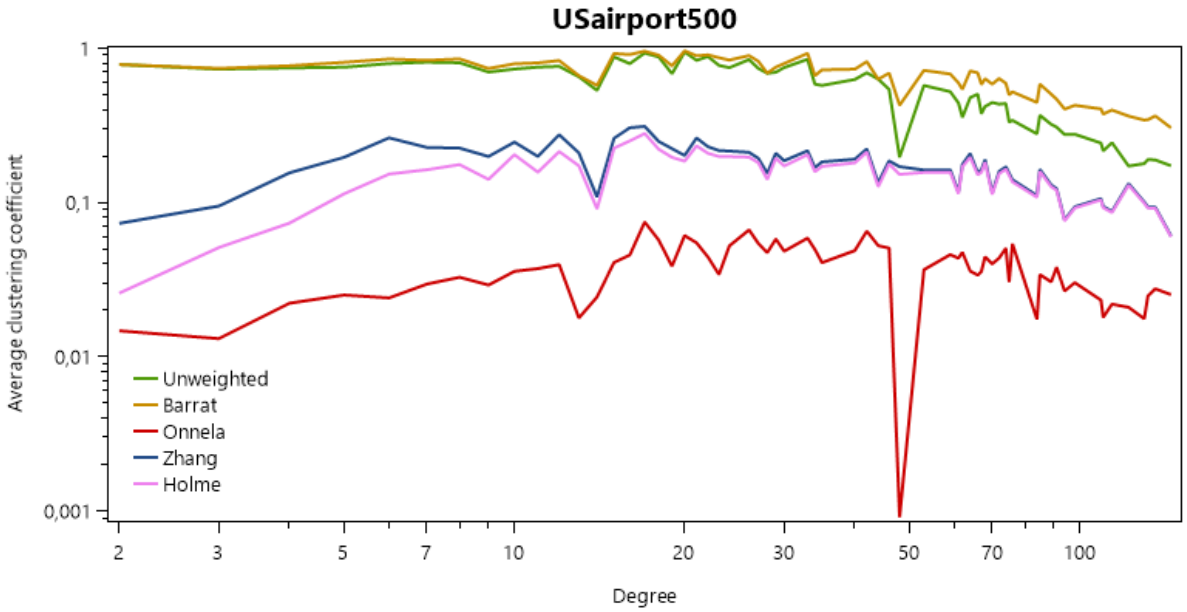
Sociální síť podobná Facebooku (zkráceně označována OClinks_w) je sociální síť zachycující komunikaci mezi jejími uživateli. Vrcholy této sítě představují jednotliví uživatelé a hrany představují zprávy mezi nimi. Ohodnocení hran představuje počet zpráv, které si dva uživatelé vyměnili. Jedná se o neorientovanou síť s 1888 vrcholy a 20296 hranami. Tento data set byl použit například v článku T. Opsahla [15].



Obrázek 20: Závislost průměrného shlukovacího koeficientu na stupni vrcholu (OClinks_w)

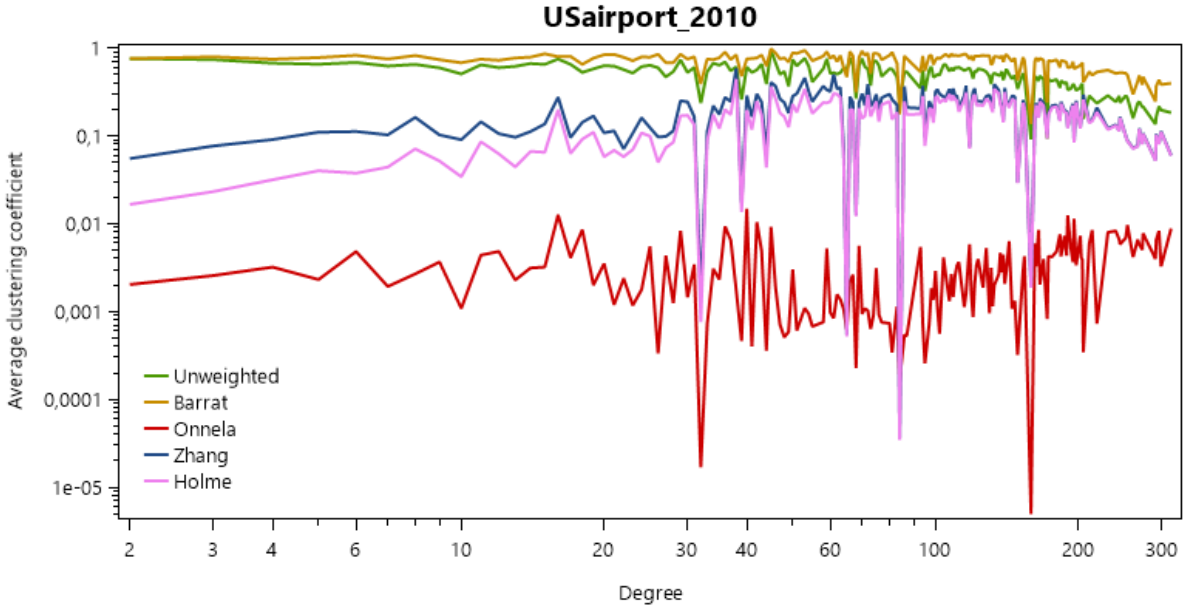
6.2.4 Sítě letišť v USA

Budeme pracovat se dvěma datasety reprezentujícími síť letišť v USA. První dataset (zkráceně označovaný USairport500) je přepravní sít mezi 500 nejvytíženějšími letišti v USA. Vrcholy této sítě představují americká letiště. Pokud mezi dvěma letišti byl naplánován let v roce 2001, existuje mezi nimi hrana. Váha této hrany pak představuje součet dostupných sedadel ve všech naplánovaných letech mezi dvěma letišti. Přesto, že je tato síť ze své podstaty orientovaná, množství míst v naplánovaných letech mezi dvěma letišti je obvykle symetrické. Tato síť je zpracována jako neorientovaná síť s 500 vrcholy a 2980 hranami. Tento dataset byl použit v článku V. Colizzy [17].



Obrázek 21: Závislost průměrného shlukovacího koeficientu na stupni vrcholu (USairport500)

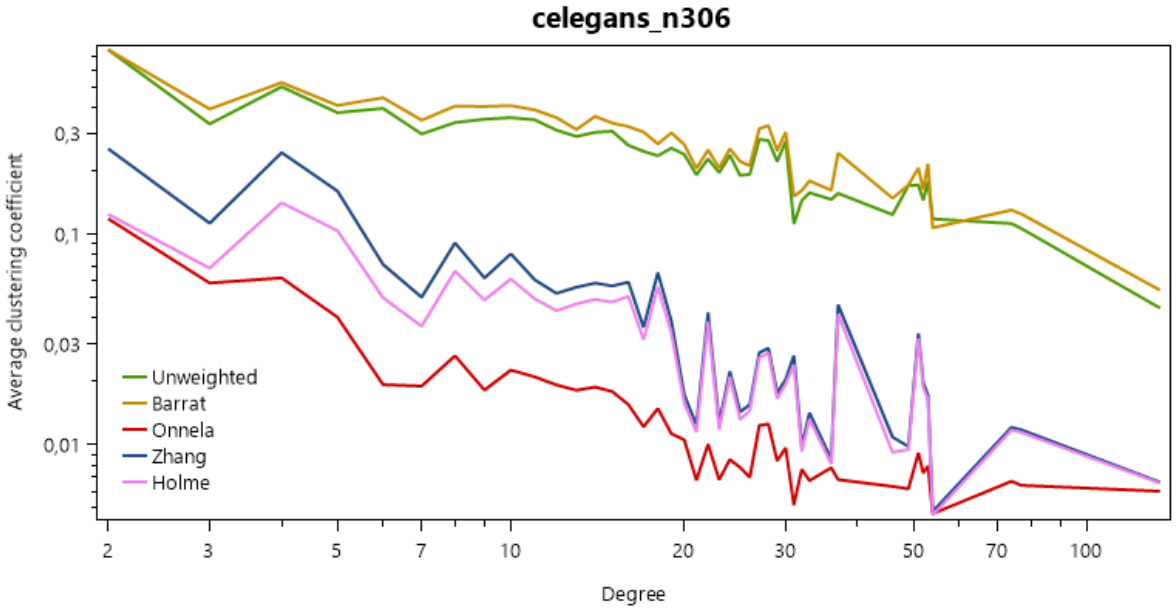
Druhý dataset (zkráceně označovaný USairport_2010) je přepravní síť mezi všemi letišti v USA v roce 2010. Vrcholy opět představují letiště a hrany jsou ohodnoceny stejným způsobem jako v prvním datasetu. Tato síť je zpracována jako neorientovaná síť s 1574 vrcholy a 14118 hranami. Tento dataset použil ve svém článku T. Opsahl [18].



Obrázek 22: Závislost průměrného shlukovacího koeficientu na stupni vrcholu (USairport_2010)

6.2.5 Neurální síť Háďátka obecného

Neurální síť Háďátka obecného (zkráceně označována celegans_n306) je první neurální síti živočicha, kterou se podařilo vědcům kompletně zmapovat. Vrcholy této sítě představují neurony. Hrany představují spojení synapsí nebo mezerovým spojem. Ohodnocení hran představuje počet těchto spojení. Jedná se neorientovanou síť s 306 vrcholy a 2345 hranami. Tímto datasetem se ve své práci zabývali Watts a Strogatz [16].



Obrázek 23: Závislost průměrného shlukovacího koeficientu na stupni vrcholu (celegans_n306)

6.2.6 Vyhodnocení

Na grafech všech datasetů napříč různými typy sítí (technologické, sociální a informační, biologické) můžeme vidět stejný trend, kdy s rostoucím stupněm vrcholu klesá průměrný shlukovací koeficient. Na grafech 17, 18, 19, 21 a 22 si můžeme všimnout, že shlukovací koeficient v sociálních-informačních a přepravních sítích je velice podobný. Na obrázcích 16, 20 a 23 můžeme vidět, že shlukovací koeficient v těchto sítích je výrazně nižší. Z grafů lze vyčíst, že C a $\tilde{C}_B$ jsou si velmi podobné. To je dáno faktem, že $\tilde{C}_B$ narozdíl od ostatních vážených variant shlukovacích koeficientů není normalizován maximální váhou hrany v sítí $max(w)$ a zohledňuje pouze lokální topologii jako C. Dále si můžeme všimnout, že $\tilde{C}_H$ a $\tilde{C}_Z$ si jsou velmi podobné. To je dáno faktem, že výpočty těchto dvou koeficientů se téměř neliší a s rostoucím stupněm vrcholu se jejich rozdíl snižuje. $\tilde{C}_O$ se od ostatních vážených variant odlišuje nejvíce, neboť vyžaduje aby váhy všech tří hran v trojúhelnících byly rovny globálnímá maximu $max(w)$. Většinou platí, že shlukovací koeficient u sociálních sítích bývá vyšší, zatím co u biologických a technologických sítích bývá nižší. Vždy však záleží na konkrétním datasetu a nelze toto tvrzení zobecnit.

7 Závěr

Cílem této práce bylo vytvořit implementace algoritmů pro výpočet shlukovacího koeficientu na ohodnocených (vážených) komplexních sítích.

V práci jsme se seznámili s komplexními sítěmi a jejich vlastnostmi. Probrali jsme si 4 základní typy sítí (technologické, sociální, informační, biologické). Uvedli jsme si jejich nejznámější zástupce. Seznámili jsme se se shlukovacím koeficientem a jeho variantami výpočtů pro ohodnocené (vážené) sítě. V C# jsme implementovali tyto výpočty a vytvořili programové komponenty, které mohou být využity jako podklad pro další práci. Dále byla vytvořena aplikace s jednoduchým uživatelským rozhraním, která umožňuje uživateli provádění výpočtů, ukládání výsledků a jejich vizualizaci. Pomocí této aplikace jsme zpracovali několik různých datasetů reprezentujících jednotlivé typy komplexních sítí. V této práci jsme se zabývali pouze neorientovanými sítěmi.

V budoucnu je možno aplikaci rozšířit o výpočty nad orientovanými sítěmi. Použitelnost implementovaných výpočtů na orientovaných sítích je potřeba otestovat a případně výpočty pro tento účel upravit.

Literatura

[1] BARABÁSI, Albert-László a Márton PÓSFAI. *Network science*. Cambridge, United Kingdom: Cambridge University Press, 2016. ISBN 978-1107076266.

[2] KOVÁŘ, Petr. *Úvod do teorie grafů* [online]. FEI VŠB - TUO, 2016 [cit. 2019-04-06]. Dostupné z: `http://homel.vsb.cz/~kov16/files/uvod_do_teorie_grafu.pdf`

[3] KOVÁŘ, Petr. *Teorie grafů*, učební text [online]. FEI VŠB - TUO, 2019 [cit. 2019-04-08]. Dostupné z: `http://homel.vsb.cz/~kov16/files/skriptum_teorie_grafu.pdf`

[4] NEWMAN, M. E. J. *Networks: an introduction*. New York: Oxford University Press, 2010. ISBN 978-0199206650.

[5] BARRAT, A., M. BARTHELEMY, R. PASTOR-SATORRAS a A. VESPIGNANI. *The architecture of complex weighted networks*. Proceedings of the National Academy of Sciences. 2004, 101(11), 3747-3752. DOI: 10.1073/pnas.0400087101. ISSN 0027-8424. Dostupné také z: `http://www.pnas.org/cgi/doi/10.1073/pnas.0400087101`

[6] ONNELA, Jukka-Pekka, Jari SARAMÄKI, János KERTÉSZ a Kimmo KASKI. *Intensity and coherence of motifs in weighted complex networks*. Physical Review E. 2005, 71(6), 3747-3752. DOI: 10.1103/PhysRevE.71.065103. ISSN 1539-3755. Dostupné také z: `https://link.aps.org/doi/10.1103/PhysRevE.71.065103`

[7] ZHANG, Bin, Steve HORVATH, János KERTÉSZ a Kimmo KASKI. *A General Framework for Weighted Gene Co-Expression Network Analysis*. Statistical Applications in Genetics and Molecular Biology. 2005, 4(1), 3747-3752. DOI: 10.2202/1544-6115.1128. ISSN 1544-6115. Dostupné také z: `https://www.degruyter.com/view/j/sagmb.2005.4.issue-1/sagmb.2005.4.1.1128/sagmb.2005.4.1.1128.xml`

[8] HOLME, Petter, Sung MIN PARK, Beom Jun KIM a Christofer R. EDLING. *Korean university life in a network perspective: Dynamics of a large affiliation network*. Physica A: Statistical Mechanics and its Applications. 2007, 373(1), 821-830. DOI: 10.1016/j.physa.2006.04.066. ISSN 03784371. Dostupné také z: `https://linkinghub.elsevier.com/retrieve/pii/S0378437106004882`

[9] SARAMÄKI, Jari, Mikko KIVELÄ, Jukka-Pekka ONNELA, Kimmo KASKI a János KERTÉSZ. *Generalizations of the clustering coefficient to weighted complex networks: Dynamics of a large affiliation network*. Physical Review E. 2007, 75(2), 821-830. DOI: 10.1103/PhysRevE.75.027105. ISSN 1539-3755. Dostupné také z: `https://link.aps.org/doi/10.1103/PhysRevE.75.027105`

[10] CLAUSET, Aaron. *Network Analysis and Modeling* [online]. Lecture 1. Santa Fe Institute, 2017 [cit. 2019-04-06]. Dostupné z: http://tuvalu.santafe.edu/~aaronc/courses/5352/csci5352_2017_L1.pdf

[11] GRANDJEAN, Martin. *La connaissance est un réseau*. Perspective sur l'organisation archivistique et encyclopédique. Les cahiers du numérique [online]. 2014, 10(3), 37-54 [cit. 2019-04-20]. DOI: 10.3166/lcn.10.3.37-54. ISSN 14693380. Dostupné z: http://lcn.revuesonline.com/article.jsp?articleId=19667

[12] MILGRAM, Stanley. *The Small-World Problem*. Psychology Today. Květen 1967, vol. 1, no. 1, s. 61-67

[13] DOBSON, Ian, Benjamin A. CARRERAS, Vickie E. LYNCH a David E. NEWMAN. *Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization*. Chaos: An Interdisciplinary Journal of Nonlinear Science [online]. 2007, 17(2) [cit. 2019-04-17]. DOI: 10.1063/1.2737822. ISSN 1054-1500. Dostupné z: http://aip.scitation.org/doi/10.1063/1.2737822

[14] NEWMAN, M. E. J. *The structure of scientific collaboration networks*. Proceedings of the National Academy of Sciences [online]. 2001, 98(2), 404-409 [cit. 2019-04-20]. DOI: 10.1073/pnas.98.2.404. ISSN 0027-8424. Dostupné z: http://www.pnas.org/cgi/doi/10.1073/pnas.98.2.404

[15] OPSAHL, Tore a Pietro PANZARASA. *Clustering in weighted networks*. Social Networks [online]. 2009, 31(2), 155-163 [cit. 2019-04-21]. DOI: 10.1016/j.socnet.2009.02.002. ISSN 03788733. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0378873309000070

[16] WATTS, Duncan J. a Steven H. STROGATZ. *Collective dynamics of 'small-world' networks*. Nature [online]. 1998, 393(6684), 440-442 [cit. 2019-04-21]. DOI: 10.1038/30918. ISSN 0028-0836. Dostupné z: urlhttp://www.nature.com/articles/30918

[17] COLIZZA, Vittoria, Romualdo PASTOR-SATORRAS a Alessandro VESPIGNANI. *Reaction–diffusion processes and metapopulation models in heterogeneous networks*. Nature Physics [online]. 2007, 3(4), 276-282 [cit. 2019-04-23]. DOI: 10.1038/nphys560. ISSN 1745-2473. Dostupné z: http://www.nature.com/articles/nphys560

[18] OPSAHL, Tore. *Why Anchorage is not (that) important: Binary ties and Sample selection* [online]. 2011 [cit. 2019-04-023]. Dostupné z: https://toreopsahl.com/2011/08/12/why-anchorage-is-not-that-important-binary-ties-and-sample-selection/

[19] NEWMAN, M. E. J. *Finding community structure in networks using the eigenvectors of matrices*. Physical Review E [online]. 2006, 74(3) [cit. 2019-04-23]. DOI: 10.1103/Phys-

RevE.74.036104. ISSN 1539-3755. Dostupné z: https://link.aps.org/doi/10.1103/PhysRevE.74.036104

[20] NELSON, Douglas L., Cathy L. MCEVOY a Thomas A. SCHREIBER. *The University of South Florida free association, rhyme, and word fragment norms*. Behavior Research Methods, Instruments, & Computers [online]. 2004, 36(3), 402-407 [cit. 2019-04-24]. DOI: 10.3758/BF03195588. ISSN 0743-3808. Dostupné z: http://www.springerlink.com/index/10.3758/BF03195588

A Výpisy zdrojového kódu

```
public static SortedList<int, double> CalculateUnweighted(Network network)
{
   SortedList<int, double> list = new SortedList<int, double>();

   foreach (var vertex in network.Vertices.Values)
   {
      double coefficient = 0;
      int degree = vertex.Degree;
      if (degree > 1)
      {
         double sum = 0;
         foreach (var n1 in vertex.Neighbours.Keys)
         {
            foreach (var n2 in vertex.Neighbours.Keys)
            {
               if (n1.Neighbours.ContainsKey(n2))
               {
                  sum++;
               }
            }
         }
         coefficient = sum / (degree * (degree - 1));
      }
      list.Add(vertex.Id, coefficient);
   }
   return list;
}
```

Výpis 3: Metoda výpočtu neváženého shlukovacího koeficientu

```
public static SortedList<int, double> CalculateBarrat(Network network)
{
   SortedList<int, double> list = new SortedList<int, double>();

   foreach (var vertex in network.Vertices.Values)
   {
      double coefficient = 0;
      int degree = vertex.Degree;
      if (degree > 1)
      {
         double sum = 0;
         foreach (var n1 in vertex.Neighbours)
         {
            foreach (var n2 in vertex.Neighbours)
            {
               if (n1.Key.Neighbours.ContainsKey(n2.Key))
               {
                  sum += (n1.Value + n2.Value) / 2;
               }
            }
         }
         coefficient = sum / (vertex.Neighbours.Values.Sum() * (degree - 1));
      }
      list.Add(vertex.Id, coefficient);
   }
   return list;
}
```

Výpis 4: Metoda výpočtu shlukovacího koeficientu podle Barrata

```
public static SortedList<int, double> CalculateOnnela(Network network)
{
   SortedList<int, double> list = new SortedList<int, double>();

   foreach (var vertex in network.Vertices.Values)
   {
      double coefficient = 0;
      int degree = vertex.Degree;
      if (degree > 1)
      {
         double sum = 0;
         foreach (var n1 in vertex.Neighbours)
         {
            foreach (var n2 in vertex.Neighbours)
            {
               if (n1.Key.Neighbours.TryGetValue(n2.Key, out double value))
               {
                  sum += Math.Pow(n1.Value * n2.Value * value, 1.0 / 3.0) /
                     network.MaxEdgeWeight;
               }
            }
         }
         coefficient = sum / (degree * (degree - 1));
      }
      list.Add(vertex.Id, coefficient);
   }
    return list;
}
```

Výpis 5: Metoda výpočtu shlukovacího koeficientu podle Onnely

```
public static SortedList<int, double> CalculateZhang(Network network)
{
   SortedList<int, double> list = new SortedList<int, double>();
   double maxEdgeWeightCubeRoot = network.MaxEdgeWeight*network.MaxEdgeWeight;

   foreach (var vertex in network.Vertices.Values)
   {
      double coefficient = 0;
      if (vertex.Degree > 1)
      {
         double sum1 = 0;
         double sum2 = 0;
         foreach (var n1 in vertex.Neighbours)
         {
            foreach (var n2 in vertex.Neighbours)
            {
               double temp = (n1.Value * n2.Value) / (maxEdgeWeightCubeRoot);
               if (n1.Key.Neighbours.TryGetValue(n2.Key, out double value))
               {
                  sum1 += temp * value / network.MaxEdgeWeight;
               }
               if (n1.Key != n2.Key)
               {
                  sum2 += temp;
                }
            }
         }
                        coefficient = sum1 / sum2;
      }
      list.Add(vertex.Id, coefficient);
   }
   return list;
}
```

Výpis 6: Metoda výpočtu shlukovacího koeficientu podle Zhanga

```
public static SortedList<int, double> CalculateHolme(Network network)
{
   SortedList<int, double> list = new SortedList<int, double>();

   foreach (var vertex in network.Vertices.Values)
   {
      double coefficient = 0;
      if (vertex.Degree > 1)
      {
         double sum1 = 0;
         double sum2 = 0;
         foreach (var n1 in vertex.Neighbours)
         {
            foreach (var n2 in vertex.Neighbours)
            {
               double temp = n1.Value * n2.Value;
               if (n1.Key.Neighbours.TryGetValue(n2.Key, out double value))
               {
                  sum1 += temp * value;
                }
                                    sum2 += temp;
            }
         }
         coefficient = sum1 / (network.MaxEdgeWeight * sum2);
      }
      list.Add(vertex.Id, coefficient);
   }
   return list;
}
```

Výpis 7: Metoda výpočtu shlukovacího koeficientu podle Holmeho

```
public static SortedList<int, double> Distribution(Network network, SortedList<
    int, double> coefficients)
{
   var sum = new SortedList<int, double>();
   var count = new SortedList<int, int>();
   var distribution = new SortedList<int, double>();

   foreach (var vertex in network.Vertices.Values)
   {
      int degree = vertex.Degree;
      if (degree < 2)
      {
         continue;
      }
      if (!sum.ContainsKey(vertex.Degree))
      {
         sum.Add(vertex.Degree, coefficients[vertex.Id]);
         count.Add(vertex.Degree, 1);
      }
      else
      {
         sum[vertex.Degree] += coefficients[vertex.Id];
         count[vertex.Degree] += 1;
      }
   }
   foreach (var k in sum.Keys)
   {
      distribution.Add(k, sum[k] / count[k]);
   }
   return distribution;
}
```

Výpis 8: Metoda pro získaní distribuce průměrného shlukovacího koeficientu vůči stupni vrcholu

B Příloha v IS EDISON

- Ve složce “Aplikace” je k dispozici spustitelná aplikace “Weighted clustering coefficient calculations.exe”.
- Ve složce “Datasety” jsou k dispozici všechny datasety použité v této práci.
- Ve složce “Výsledky výpočtů” jsou vypočítáné výsledky pro všechny datasety prostřednictvím přiložené aplikace. Výsledky každého datasetu se skládají ze tří souborů: CSV s lokálními shlukovacími koeficienty, CSV s distribucí průměného shlukovacího koeficientu vůči stupni vrcholu a PNG s grafem distribuce průměného shlukovacího koeficientu vůči stupni vrcholu.
- Ve složce “Zdrojový kód” je k dispozici řešení pro Visual Studio obsahující veškeré zdrojové kódy.